#5

500.41226X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s):	MATAUBAYASHI, et al.
Serial No.:	Not assigned
Filed:	February 25, 2002
Title:	METHOD OF SEARCHING SIMILAR DOCUMENT, SYSTEM FOR PERFORMING THE SAME AND PROGRAM FOR PROCESSING THE SAME
Group:	Not assigned

LETTER CLAIMING RIGHT OF PRIORITY

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231

February 25, 2002

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Application No.(s) 2001-128934 filed April 26, 2001.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl. I. Brundidge
Registration No. 29,621

CIB/amr
Attachment
(703) 312-6600

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application:　２００１年　４月２６日

出願番号
Application Number:　特願２００１－１２８９３４

[ST.10/C]:　[ＪＰ２００１－１２８９３４]

出願人
Applicant(s):　株式会社日立製作所

２００２年　１月２９日

特許庁長官
Commissioner,
Japan Patent Offic　及川耕造

出証番号　出証特２００２－３００２３２６

【書類名】　特許願

【整理番号】　K01003091

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/30

【発明者】

【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内

【氏名】　松林　忠孝

【発明者】

【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内

【氏名】　多田　勝己

【発明者】

【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内

【氏名】　里　佳史

【発明者】

【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内

【氏名】　稲場　靖彦

【発明者】

【住所又は居所】　神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内

【氏名】　野田　十悟

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社日立製作所

【代理人】

【識別番号】　100083552

【弁理士】

【氏名又は名称】　秋田　収喜

【電話番号】　03-3893-6221

【手数料の表示】

【予納台帳番号】　014579

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　類似文書検索方法及びその実施システム並びにその処理プログラム

【特許請求の範囲】

【請求項１】　指定された文書と類似する文書を検索する類似文書検索方法において、

所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出するステップと、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出するステップと、

前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出するステップと、前記算出された類似度算出結果を検索結果として出力するステップとを有することを特徴とする類似文書検索方法。

【請求項２】　前記抽出された特徴語候補に対応する特徴語にその構成特徴語を示す構成特徴語情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定することを特徴とする請求項１に記載された類似文書検索方法。

【請求項３】　前記抽出された種文書の構成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出するステップを有し、前記重み係数を乗じた類似度を算出することを特徴とする請求項１または請求項２のいずれかに記載された類似文書検索方法。

【請求項４】　指定された文書と類似する文書を検索する類似文書検索システムにおいて、

所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出する文書解析処理部と、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出する特徴語抽出処理部と、

前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出する種文書類似度算出処理部と、前記算出された類似度算出結果を検索結果として出

力する検索結果出力処理部とを備えることを特徴とする類似文書検索システム。

【請求項５】　前記抽出された特徴語候補に対応する特徴語にその構成特徴語を示す構成特徴語情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定する複合特徴語判定処理部を備えることを特徴とする請求項４に記載された類似文書検索システム。

【請求項６】　前記抽出された種文書の構成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出する重み係数算出処理部を備え、前記重み係数を乗じた類似度を算出することを特徴とする請求項４または請求項５のいずれかに記載された類似文書検索システム。

【請求項７】　指定された文書と類似する文書を検索する類似文書検索システムとしてコンピュータを機能させる為のプログラムにおいて、

所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出する文書解析処理部と、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出する特徴語抽出処理部と、

前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出する種文書類似度算出処理部と、前記算出された類似度算出結果を検索結果として出力する検索結果出力処理部としてコンピュータを機能させることを特徴とするプログラム。

【請求項８】　前記抽出された特徴語候補に対応する特徴語にその構成特徴語を示す構成特徴語情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定する複合特徴語判定処理部としてコンピュータを機能させることを特徴とする請求項７に記載されたプログラム。

【請求項９】　前記抽出された種文書の構成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出する重み係数算出処理部としてコンピュータを機能させることを特徴とする請求項７または請求項８のいずれかに記載されたプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は指定された文書と類似する文書を検索する類似文書検索システムに関し、特にユーザから指定された文書に記述されている特徴語を含む文書を類似文書として文書データベースの中から検索する類似文書検索システムに適用して有効な技術に関するものである。

【０００２】

【従来の技術】

近年、組織内での業務の効率化や、業務の質を向上させる為に、組織内の個人の知識を共有し、再利用することを目的とする知識管理システムへの要求が高まってきている。

【０００３】

特に企業内で活用する知識管理システムに対しては、有識者の経験やノウハウ等を文書化し、知識として共有、活用することへの要望が高まっており、組織内で非定型に蓄えられた大量の知識の中から、ユーザが所望するものを簡単にかつ適切に取得する高精度な検索機能が重要になってきている。

【０００４】

この様な要求に応える技術として、ユーザが自分の所望する内容を含んだ文書（以下、種文書と呼ぶ）を例示し、その文書と類似する文書を検索する類似文書検索技術が注目されている。

【０００５】

類似文書検索の方法としては、例えば、文書内に出現する単語（以下、特徴語と呼ぶ）の出現頻度を要素とするベクトル（以下、特徴ベクトルと呼ぶ）を用いて文書間の類似度を算出する技術（以下、従来技術１と呼ぶ）が、“Information Retrieval”（William B.Frakes、Prentice Hall PTR、pp.363～376）に開示されている。

【０００６】

従来技術１の概要は次の通りである。文書データベースに文書を登録する際に、登録対象となる文書中に含まれる特徴語の出現頻度を登録文書の特徴ベクトル

（以下、登録文書特徴ベクトルと呼ぶ）として作成しておく。

【０００７】

類似文書の検索時は、検索条件として指定された種文書の特徴ベクトル（以下、種文書特徴ベクトルと呼ぶ）と各登録文書特徴ベクトルとのベクトル空間内においてなす角度の余弦を、文書間の類似度として算出する。

【０００８】

図２０は従来技術１の処理手順の一例を示す図である。以下、従来技術１の処理手順を図２０のＰＡＤ(Problem Analysis Diagram) 図を用いて説明する。

【０００９】

まずステップ２００において、文書の登録処理か類似文書の検索処理かを判定する。そして、文書の登録処理と判定された場合には登録文書特徴ベクトル生成ステップ２１０を実行し、登録対象文書に対する登録文書特徴ベクトルを生成する。

【００１０】

また、ステップ２００において類似文書の検索処理と判定された場合には、種文書特徴ベクトル生成ステップ２２０を実行し、検索条件として指定された種文書に対する種文書特徴ベクトルを作成する。

【００１１】

次にステップ２２１を実行し、全登録文書に対して類似度算出ステップ２２２を繰り返し実行する。類似度算出ステップ２２２では、前記種文書特徴ベクトルと登録文書特徴ベクトルが、ベクトル空間内においてなす角度の余弦を文書間の類似度として算出する。

【００１２】

図２１は従来技術１における特徴ベクトル生成処理の一例を示す図である。以下、図２０に示した登録文書特徴ベクトル生成ステップ２１０及び種文書特徴ベクトル生成ステップ２２０として実行される従来技術１における特徴ベクトル生成処理について、図２１に示したＰＡＤ図を用いて説明する。

【００１３】

特徴ベクトル生成処理では、まずステップ３０１において、特徴ベクトルの生

成処理対象となる文書を読み込む。次にステップ３０２において、上記ステップ３０１で読み込まれた処理対象文書から特徴語を抽出する。

【００１４】

そしてステップ３０３において、上記ステップ３０２で抽出された各特徴語の出現頻度を計数する。最後にステップ３０４において、上記ステップ３０２で抽出された各特徴語と、上記ステップ３０３で計数した各特徴語の出現頻度を特徴ベクトルの要素として格納する。以上が、従来技術１の処理手順である。

【００１５】

図２２は従来技術１の概要を示す図である。以下、図２２を用いて従来技術１の処理例を説明する。

【００１６】

従来技術１では、まず処理要求判定ステップ４１０において、入力された処理要求が登録処理であるか、或いは検索処理であるかを判定する。そして、入力された処理要求が登録処理である場合には、ステップ２１０が実行される。

【００１７】

ステップ２１０では、登録用文書１及び文書２中に含まれる特徴語を抽出すると共に各文書内での出現頻度を計数し、各文書に対応する登録文書特徴ベクトル４０３及び４０４を生成する。

【００１８】

ここで、登録文書特徴ベクトル４０３“文書１（”ＬＡＮ”、１）（”構築”、１）…”は、「文書１」の特徴ベクトルであり、特徴語“ＬＡＮ”が１回、特徴語“構築”が１回出現していることを表している。

【００１９】

また、前記処理要求判定ステップ４１０で類似文書の検索処理と判定された場合には、検索条件で指定された種文書４０６から特徴語を抽出し、ステップ２２０で該種文書に対応する種文書特徴ベクトル４０７を生成する。

【００２０】

次に、種文書特徴ベクトル４０７と前記ステップ２１０で生成された各登録文書の登録文書特徴ベクトルとのなす角の余弦を類似度として算出する。

【００２１】

一般に、２つのベクトルＡ及びベクトルＢのなす角の余弦は、数１の様に算出される。ここで“Ａ・Ｂ”は、ベクトルＡとベクトルＢの内積を表し、“|Ａ|”は、ベクトルＡの大きさを表す。

【００２２】

【数１】

数1

ベクトルAとベクトルBのなす角度の余弦

$$= \frac{A \cdot B}{|A| \times |B|}$$

【００２３】

図２２に示した種文書特徴ベクトル４０７と登録文書特徴ベクトル４０３及び登録文書特徴ベクトル４０４のなす角の余弦は、ベクトルＡを種文書特徴ベクトル４０７、ベクトルＢを登録文書特徴ベクトル４０３または登録文書特徴ベクトル４０４として、それぞれ数２、数３の様に算出される。

【００２４】

【数２】

数2

$$\frac{1\times0+1\times0+1\times1+1\times0}{\sqrt{1^2+1^2+1^2+1^2} \times \sqrt{1^2+1^2+1^2+1^2+1^2+1^2}}$$

$$= \frac{1}{2\sqrt{6}} = 0.204$$

【００２５】

【数３】

数3

$$\frac{1\times1+1\times1+1\times1+1\times0}{\sqrt{1^2+1^2+1^2+1^2} \times \sqrt{1^2+1^2+1^2+1^2+1^2}}$$

$$= \frac{3}{2\sqrt{5}} = 0.670$$

【００２６】

この結果として、種文書に対する各登録文書の類似度算出結果４０８が出力される。以上が、従来技術１の処理例である。

【００２７】

以上説明した様に従来技術１によれば、登録文書中に含まれる特徴語を抽出した登録文書特徴ベクトルを予め作成しておき、検索条件として指定された種文書に対応する種文書特徴ベクトルとの余弦を類似度として算出することで、文書データベース中から内容の類似する文書を検索することができる。

【００２８】

【発明が解決しようとする課題】

しかし従来技術１では、特徴ベクトルの要素である特徴語が複数の単語で構成されている場合に、検索漏れが発生するという問題がある。

【００２９】

図２３は従来技術１の問題点を示す図である。以下、図２３を用いて、従来技術１の問題点を説明する。本図では、文書３「地図情報閲覧ソフトを開発、発売したＡ社は、・・・」及び文書４「多くの地図閲覧ソフトが発売されているが、・・・」が登録された文書データベースに対して、種文書「最新の地図閲覧ソフトについて」が入力された場合の例を表している。

【００３０】

まず文書の登録処理として、ステップ２１０が実行され、各文書に対応する登録文書特徴ベクトル４０３ａ及び４０４ａが生成される。本図に示した例では、文書３に対応する特徴ベクトル４０３ａとして“文書３（”地図”、1）（”閲覧”、1）（”ソフト”、1）（”発売”、1）”が生成され、文書４に対応する特徴ベクトル４０４ａとして“文書４（”地図閲覧ソフト”、1）（”発売”、1）”が生成される。

【００３１】

次に類似文書の検索処理として、種文書特徴ベクトル生成処理ステップ２２０が実行され、種文書に対応する種文書特徴ベクトル４０７ａが生成される。本図に示した例では、種文書特徴ベクトル４０７ａとして、“種文書（”地図閲覧ソ

フト”、1）”が生成される。

【００３２】

そして類似度算出ステップ２２２において、種文書に対する各登録文書の類似度を算出する。この結果、類似度算出結果４０８ａが出力される。本図に示した例では、数４及び数５に示す様に、文書３の類似度０．０００及び文書４の類似度０．７１０と算出される。

【００３３】

【数４】

数4

$$\frac{1\times 0}{\sqrt{1^2}\times\sqrt{1^2+1^2+1^2+1^2}} = \frac{0}{2} = 0.000$$

【００３４】

【数５】

数5

$$\frac{1\times 1}{\sqrt{1^2}\times\sqrt{1^2+1^2}} = \frac{1}{\sqrt{2}} = 0.710$$

【００３５】

この結果、文書３の内容は種文書に対して関連があるにも関わらず、従来技術１では文書３の内容は種文書に対して全く類似していないものと算出されてしまう。

【００３６】

これは、種文書の特徴ベクトルの要素として抽出される特徴語が複数の単語で構成されているにもかかわらず、最長一致の特徴語「地図閲覧ソフト」のみを特徴ベクトルの要素として類似度算出に用いた為に、特徴語を構成する各単語の持つ個々の概念が類似度に反映されないことによるものである。すなわち、特徴語

を構成する各単語それぞれを含む登録文書に対して類似度が付与されず、検索漏れが発生してしまうことになる。

【００３７】

一方、前記の最長一致の特徴語「地図閲覧ソフト」の代わりに、「地図閲覧ソフト」を構成する各単語「地図」「閲覧」「ソフト」を用いることで前記の様な検索漏れを防止することができるが、この場合には「地図閲覧ソフト」とは類似度の低い文書がノイズとして検索される可能性が高くなる。以上が従来技術１の問題点である。

【００３８】

本発明の目的は上記問題を解決し、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能な技術を提供することにある。本発明の他の目的は検索漏れが無くノイズの少ない高精度な類似文書検索を実現することが可能な技術を提供することにある。

【００３９】

【課題を解決するための手段】

本発明は、指定された文書と類似する文書を検索する類似文書検索システムにおいて、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するものである。

【００４０】

本発明の類似文書検索システムでは、処理対象文書から抽出された特徴語候補が複数の特徴語から構成されている複合特徴語であるか、単一の単語から構成されている単独特徴語であるかを判定し、複合特徴語と判定された場合には複合特徴語及びその複合特徴語を構成する構成特徴語を特徴語として抽出し、単独特徴語と判定された場合には該特徴語そのものを抽出する。

【００４１】

すなわち、複合特徴語とその複合特徴語を構成する構成特徴語を抽出し、その抽出した複合特徴語及び構成特徴語を類似度算出に使用することにより、検索漏れの無い高精度な類似文書検索を実現することが可能となる。

【００４２】

以上の様に本発明の類似文書検索システムによれば、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するので、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能である。

【００４３】

【発明の実施の形態】

（実施形態１）

以下に指定された種文書中の複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索する実施形態１の類似文書検索システムについて説明する。

【００４４】

図１は本実施形態の類似文書検索システムの概略構成を示す図である。図１に示す様に本実施形態の類似文書検索システムは、システム制御処理部１１０と、登録制御処理部１１１と、検索制御処理部１１２と、登録文書取得処理部１２０と、登録文書特徴ベクトル登録処理部１２１と、検索条件解析処理部１３０と、種文書類似度算出処理部１３１と、検索結果出力処理部１３２と、登録文書特徴ベクトル読込処理部１６０と、類似度算出処理部１６１と、特徴ベクトル生成処理部１７０と、特徴語抽出処理部１７１と、文書解析処理部１７２と、複合特徴語判定処理部１７３と、出現頻度計数処理部１７４とを有している。

【００４５】

システム制御処理部１１０は、キーボード１０１から入力されたコマンドを解析し、登録制御処理部１１１または検索制御処理部１１２を起動する処理部である。登録制御処理部１１１は、登録文書取得処理部１２０を起動し、登録対象として指定された文書の特徴ベクトルの磁気ディスク装置１０３への格納を制御する処理部である。

【００４６】

検索制御処理部１１２は、検索条件解析処理部１３０、種文書類似度算出処理部１３１、検索結果出力処理部１３２を起動し、検索条件で指定された種文書に類似する文書の検索を制御する処理部である。

【００４７】

登録文書取得処理部１２０は、登録対象の文書を取得する処理部である。登録文書特徴ベクトル登録処理部１２１は、登録対象の文書の特徴ベクトルを磁気ディスク装置１０３へ格納する処理部である。検索条件解析処理部１３０は、検索条件で指定された種文書を取得する処理部である。

【００４８】

種文書類似度算出処理部１３１は、登録文書特徴ベクトル読込処理部１６０及び類似度算出処理部１６１を起動し、種文書から抽出された特徴語と各登録文書との間の類似度を算出する処理部である。検索結果出力処理部１３２は、前記算出された類似度算出結果を検索結果として出力する処理部である。

【００４９】

登録文書特徴ベクトル読込処理部１６０は、磁気ディスク装置１０３に格納された登録文書特徴ベクトルファイル１８０を読み込む処理部である。類似度算出処理部１６１は、種文書特徴ベクトルに対する登録文書特徴ベクトルのなす角度の余弦を算出し、種文書に対する登録文書の類似度を算出する処理部である。

【００５０】

特徴ベクトル生成処理部１７０は、特徴語抽出処理部１７１及び出現頻度計数処理部１７４を起動し、処理対象文書の特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該処理対象文書の特徴語として抽出して処理対象文書の特徴ベクトルを生成する処理部である。

【００５１】

特徴語抽出処理部１７１は、文書解析処理部１７２及び複合特徴語判定処理部１７３を起動し、処理対象文書から特徴語または複合特徴語及び構成特徴語を抽出する処理部である。文書解析処理部１７２は、登録対象の文書である登録文書や所望の検索内容を含んだ種文書等の処理対象文書から特徴語の候補となる特徴語候補を抽出する処理部である。

【００５２】

複合特徴語判定処理部１７３は、前記抽出された特徴語候補に対応する特徴語

にその構成特徴語を示す構成特徴語情報としてそれらの構成特徴語のポインタ情報が登録されている場合に、当該特徴語候補が複合特徴語であると判定する処理部である。出現頻度計数処理部１７４は、処理対象文書から抽出された各特徴語の当該処理対象文書における出現頻度を計数する処理部である。

【００５３】

類似文書検索システムをシステム制御処理部１１０、登録制御処理部１１１、検索制御処理部１１２、登録文書取得処理部１２０、登録文書特徴ベクトル登録処理部１２１、検索条件解析処理部１３０、種文書類似度算出処理部１３１、検索結果出力処理部１３２、登録文書特徴ベクトル読込処理部１６０、類似度算出処理部１６１、特徴ベクトル生成処理部１７０、特徴語抽出処理部１７１、文書解析処理部１７２、複合特徴語判定処理部１７３及び出現頻度計数処理部１７４として機能させる為のプログラムは、ＣＤ－ＲＯＭ等の記録媒体に記録され磁気ディスク等に格納された後、メモリにロードされて実行されるものとする。なお前記プログラムを記録する記録媒体はＣＤ－ＲＯＭ以外の他の記録媒体でも良い。また前記プログラムを当該記録媒体から情報処理装置にインストールして使用しても良いし、ネットワークを通じて当該記録媒体にアクセスして前記プログラムを使用するものとしても良い。

【００５４】

本実施形態の類似文書検索システムは、ディスプレイ１００、キーボード１０１、中央演算処理装置であるＣＰＵ１０２、磁気ディスク装置１０３、フロッピディスクドライブであるＦＤＤ１０４、主メモリ１０５、これらを結ぶバス１０６及び他の機器と本システムを接続するネットワーク１０８から構成される。

【００５５】

磁気ディスク装置１０３は二次記憶装置の一つであり、登録文書特徴ベクトルファイル１８０及び特徴語辞書ファイル１８１が格納される。ＦＤＤ１０４を介してフロッピディスク１０７に格納されている情報が、主メモリ１０５或いは磁気ディスク装置１０３へ読み込まれる。

【００５６】

主メモリ１０５には、システム制御処理部１１０、登録制御処理部１１１、検

索制御処理部１１２、登録文書取得処理部１２０、登録文書特徴ベクトル登録処理部１２１、検索条件解析処理部１３０、種文書類似度算出処理部１３１、検索結果出力処理部１３２及び共有ライブラリ１４０が格納されると共にワークエリア１４１が確保される。共有ライブラリ１４０には、特徴ベクトル生成処理部１７０、特徴語抽出処理部１７１及び出現頻度計数処理部１７４が格納される。

【００５７】

種文書類似度算出処理部１３１は、登録文書特徴ベクトル読込処理部１６０及び類似度算出処理部１６１で構成される。特徴ベクトル生成処理部１７０は、特徴語抽出処理部１７１及び出現頻度計数処理部１７４を呼び出す構成をとる。特徴語抽出処理部１７１は、文書解析処理部１７２及び複合特徴語判定処理部１７３で構成される。

【００５８】

登録制御処理部１１１及び検索制御処理部１１２は、キーボード１０１からのユーザによる指示に応じてシステム制御処理部１１０によって起動され、それぞれ登録文書取得処理部１２０、特徴ベクトル生成処理部１７０及び登録文書特徴ベクトル登録処理部１２１の制御と、検索条件解析処理部１３０、特徴ベクトル生成処理部１７０、種文書類似度算出処理部１３１及び検索結果出力処理部１３２の制御を行なう。

【００５９】

なお本実施形態では、キーボード１０１から入力されたコマンドにより、登録制御処理部１１１や検索制御処理部１１２が起動されるものとしたが、他の入力装置を介して入力されたコマンド或いはイベントにより起動されるものであっても構わない。

【００６０】

また、本実施形態の類似文書検索システムをこれらの処理部として機能させる為のプログラムは、磁気ディスク装置１０３、フロッピディスク１０７、或いはＭＯ、ＣＤ－ＲＯＭ、ＤＶＤ等の記録媒体（図１には示していない）に格納され、駆動装置を介して主メモリ１０５に読み込まれ、ＣＰＵ１０２によって実行されるものとするが、これらのプログラムをネットワーク１０８を介して主メモリ

１０５に読み込み、ＣＰＵ１０２によって実行することも同様に可能である。

【００６１】

更に、本実施形態では登録文書特徴ベクトルファイル１８０及び特徴語辞書ファイル１８１を磁気ディスク装置１０３に格納するものとしたが、フロッピディスク１０７、ＭＯ、ＣＤ－ＲＯＭ、ＤＶＤ等の記録媒体（図１には示していない）に格納し、駆動装置を介して主メモリ１０５に読み込み利用することも可能である。また、これらのファイルはネットワーク１０８を介して、他のシステムに接続された記録媒体（図１には示していない）に格納されるものとしても良いし、或いはネットワーク１０８に直接接続された記録媒体に格納されるものとしても構わない。

【００６２】

以下、本実施形態における類似文書検索システムの処理手順について説明する。

図本実施形態のシステム制御処理部１１０の処理内容を示す図である。まず、システム制御処理部１１０の処理手順について図２のＰＡＤ図を用いて説明する。

【００６３】

システム制御処理部１１０は、まずステップ８００で、キーボード１０１から入力されたコマンドを解析する。そしてステップ８０１で、この結果が登録実行のコマンドであると解析された場合には、ステップ８０２で登録制御処理部１１１を起動して文書の登録を行なう。またステップ８０１で、検索実行のコマンドであると解析された場合には、ステップ８０３で検索制御処理部１１２を起動して、類似文書の検索を行なう。以上が、システム制御処理部１１０の処理手順である。

【００６４】

図３は本実施形態の登録制御処理部１１１の処理内容を示す図である。図２に示したシステム制御処理部１１０のステップ８０２で起動される登録制御処理部１１１の処理手順について、図３のＰＡＤ図を用いて説明する。

【００６５】

登録制御処理部１１１では、まずステップ９００において登録文書取得処理部１２０を起動し、登録対象として指定された文書（以下、登録対象文書と呼ぶ）を読み込み、ワークエリア１４１に格納する。

【００６６】

次に、ステップ９０１において、共有ライブラリ１４０に格納されている特徴ベクトル生成処理部１７０を起動し、登録対象文書に対する特徴ベクトルを生成し、ワークエリア１４１に格納する。

【００６７】

そして、ステップ９０２において、登録文書特徴ベクトル登録処理部１２１を起動し、ワークエリア１４１に格納されている登録文書特徴ベクトルを磁気ディスク装置１０３へ格納する。以上が、登録制御処理部１１１の処理手順である。

【００６８】

図４は本実施形態の特徴ベクトル生成処理部１７０の処理内容を示す図である。図３に示した登録制御処理部１１１のステップ９０１で起動される特徴ベクトル生成処理部１７０の処理手順について、図４のＰＡＤ図を用いて説明する。

【００６９】

特徴ベクトル生成処理部１７０では、まずステップ１０００において特徴語抽出処理部１７１を起動し、ワークエリア１４１に格納された処理対象文書から特徴語を抽出する。次に、ステップ１００１において、出現頻度計数処理部１７４を起動し、ワークエリア１４１に格納された各特徴語の処理対象文書における出現頻度を計数する。以上が、特徴ベクトル生成処理部１７０の処理手順である。なお、本特徴ベクトル生成処理部１７０は共有ライブラリ１４０に格納されており、後述する文書検索処理における検索制御処理部１１２からも実行され、種文書に対する特徴ベクトルの生成においても使用される。

【００７０】

図５は本実施形態の特徴語抽出処理部１７１の処理内容を示す図である。図４に示した特徴ベクトル生成処理部１７０のステップ１０００で起動される特徴語抽出処理部１７１の処理手順について、図５のＰＡＤ図を用いて説明する。

【００７１】

特徴語抽出処理部１７１は、まずステップ１４００において文書解析処理部１７２を起動し、ワークエリア１４１に格納された処理対象文書中の文字列と特徴語辞書ファイル１８１中の特徴語とを比較し、特徴語辞書ファイル１８１中の特徴語と一致する文字列を特徴語候補として処理対象文書から抽出する。

【００７２】

次にステップ１４０１において、複合特徴語判定処理部１７３を起動し、上記ステップ１４００において抽出された特徴語候補に対応する特徴語辞書ファイル１８１中の特徴語に構成特徴語の格納位置を示すポインタ情報が登録されているかどうかを調べ、特徴語辞書ファイル１８１中の特徴語に前記ポインタ情報が登録されている場合には、当該特徴語候補が複合特徴語であると判定する。

【００７３】

そして、特徴語候補が複合特徴語であると判定された場合には、ステップ１４０２を実行し、前記ポインタ情報で示された特徴語をその複合特徴語の構成特徴語として読み出して、それらの複合特徴語及び構成特徴語を処理対象文書の特徴語として抽出する。

【００７４】

前記の様に本実施形態では、抽出された特徴語候補に対応する特徴語辞書ファイル１８１中の特徴語に構成特徴語の格納位置を示すポインタ情報が登録されているかどうかを調べることにより、当該特徴語候補が複合特徴語であるかを判定し、前記ポインタ情報を用いて構成特徴語の読み出しを行なうので、特徴語候補が複合特徴語であるかの判定及び構成特徴語の読み出しを高速に行なうことが可能である。なお、本実施形態では複合特徴語に構成特徴語を示すポインタ情報を格納するものとしたが、複合特徴語内に分割位置を格納しておくものとしても良いし、構成特徴語そのものを格納しておくものとしても良い。

【００７５】

また、ステップ１４０１において、特徴語候補が複合特徴語でないと判定された場合にはステップ１４０３を実行し、特徴語候補そのものを処理対象文書の特徴語として抽出する。以上が、特徴語抽出処理部１７１の処理手順である。

【００７６】

図６は本実施形態の検索制御処理部１１２の処理内容を示す図である。図２に示したシステム制御処理部１１０のステップ８０３で起動される検索制御処理部１１２の処理手順について、図６のＰＡＤ図を用いて説明する。

【００７７】

検索制御処理部１１２は、まずステップ１１００において、検索条件解析処理部１３０を起動し、検索条件で指定された種文書を取得する。そしてステップ１１０１において、共有ライブラリ１４０に格納された特徴ベクトル生成処理部１７０を起動し、上記ステップ１１００で取得された種文書に対する種文書特徴ベクトルを生成する。

【００７８】

次にステップ１１０２において、種文書類似度算出処理部１３１を起動し、種文書に対する各登録文書の類似度を算出する。そしてステップ１１０３において、検索結果出力処理部１３２を起動し、上記ステップ１１０１で算出された類似度算出結果を検索結果として出力する。

【００７９】

ここで、検索結果の出力先は、ディスプレイ１００に表示するものとしても良いし、ワークエリア１４１や磁気ディスク装置１０３上に格納するものとしても良い。また、類似度算出結果をディスプレイ１００に出力する場合には、類似度の降順に出力するものとしても良いし、文書に付与された管理番号の昇順或いは降順に出力するものとしても良い。以上が検索制御処理部１１２の処理手順である。

【００８０】

図７は本実施形態の種文書類似度算出処理部１３１の処理内容を示す図である。図６に示した検索制御処理部１１２のステップ１１０２で起動される種文書類似度算出処理部１３１の処理手順について、図７のＰＡＤ図を用いて説明する。

【００８１】

種文書類似度算出処理部１３１は、まずステップ１３００において、登録文書特徴ベクトル読込処理部１６０を起動し、磁気ディスク装置１０３に格納された登録文書特徴ベクトルファイル１８０を読み込み、ワークエリア１４１に格納す

る。

【００８２】

そしてステップ１３０１において、ワークエリア１４１に格納された全ての登録文書特徴ベクトルに対して、ステップ１３０２を繰り返し実行する。ステップ１３０２では、類似度算出処理部１６１を起動し、種文書特徴ベクトルに対する登録文書特徴ベクトルのなす角度の余弦を算出し、種文書に対する登録文書の類似度としてワークエリア１４１に格納する。以上が種文書類似度算出処理部１３１の処理手順である。

【００８３】

以下、本実施形態における類似文書検索システムの具体的な処理手順を図８～図１１を用いて説明する。まず、本実施形態における類似文書検索システムにおける文書の登録処理について、図８を用いて説明する。

【００８４】

図８は本実施形態の文書の登録処理の処理内容を示す図である。図８では、文書３「地図情報閲覧ソフトを開発、発売したＡ社は、・・・」及び文書４「多くの地図閲覧ソフトが発売されているが、・・・」が文書データベースに登録される場合の処理の流れを表している。

【００８５】

まず、本実施形態の類似文書検索システムにおいて、登録文書取得処理部１２０は、登録対象の文書３及び文書４を読み込み、ワークエリア１４１に格納する。次に特徴ベクトル生成処理部１７０は、登録対象の文書３及び文書４に対応する登録文書特徴ベクトル４０３ａ及び４０４ｂを作成し、ワークエリア１４１に格納する。そして、登録文書特徴ベクトル登録処理部１２１は、ワークエリア１４１上の登録文書特徴ベクトルを登録文書特徴ベクトルファイル１８０に格納する。以上が、本実施形態に示した類似文書検索システムにおける文書の登録処理である。

【００８６】

次に、本実施形態における類似文書検索システムにおける類似文書の検索処理について、図９を用いて説明する。

【００８７】

図９は本実施形態の類似文書の検索処理の処理内容を示す図である。図９では、種文書「最新の地図閲覧ソフトについて」が入力された場合の例を表している。まず、検索条件解析処理部１３０は、検索条件で指定された種文書を取得し、ワークエリア１４１に格納する。

【００８８】

そして、特徴ベクトル生成処理部１７０は、ワークエリア１４１に格納された種文書に対応する種文書特徴ベクトル４０７ｂを生成し、ワークエリア１４１に格納する。

【００８９】

次に、登録文書特徴ベクトル読込処理部１６０は、前記文書の登録処理で作成された登録文書特徴ベクトルファイル１８０を読み込み、登録文書特徴ベクトル４０３ａ及び４０４ｂをワークエリア１４１に格納する。

【００９０】

【数６】

数6

$$\frac{1\times0+1\times1+1\times1+1\times1}{\sqrt{1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2}}$$

$$=\frac{3}{2\times 2}=0.750$$

【００９１】

【数７】

数7

$$\frac{1\times1+1\times1+1\times1+1\times1}{\sqrt{1^2+1^2+1^2+1^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2}}$$

$$=\frac{4}{2\sqrt{5}}=0.894$$

【００９２】

そして、類似度算出処理部１６１は、前記ステップ１７０で生成された種文書

特徴ベクトル４０７ｂと登録文書特徴ベクトル４０３ａ及び４０４ｂのなす角度の余弦を数６及び数７の様に算出し、種文書に対する登録文書の類似度算出結果４０８ｂを出力する。以上が、本実施形態における類似文書検索システムにおける類似文書の検索処理手順である。

【００９３】

次に、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順について図１０を用いて説明する。

【００９４】

図１０は本実施形態の特徴ベクトルの生成処理の処理内容を示す図である。図１０では、種文書「最新の地図閲覧ソフトについて」が入力された場合にその特徴ベクトルが作成される手順を表している。

【００９５】

まず、文書解析処理部１７２は、ワークエリア１４１に格納された処理対象文書である種文書１６０１“最新の地図閲覧ソフトについて”中の文字列と特徴語辞書ファイル１８１中の特徴語とを比較し、特徴語辞書ファイル１８１中の特徴語と一致する文字列“地図閲覧ソフト”を特徴語候補１６０２として種文書１６０１から抽出する。

【００９６】

そして、複合特徴語判定処理部１７３は、特徴語辞書ファイル１８１中の特徴語“地図閲覧ソフト”に構成特徴語の格納位置を示すポインタ情報が登録されているかどうかを調べ、特徴語候補１６０２“地図閲覧ソフト”が複数の特徴語で構成される複合特徴語であるかを判定する。この結果、特徴語候補１６０２“地図閲覧ソフト”は複数の特徴語“地図”、“閲覧”、“ソフト”から構成されるものと判定され、複合特徴語と判定される。

【００９７】

次に、特徴語抽出処理部１７１は、上記複合特徴語判定処理部１７３の結果、複合特徴語と判定された“地図閲覧ソフト”から、これを構成する特徴語１６０４“地図”“閲覧”“ソフト”を前記ポインタ情報により抽出する。そして、出現頻度計数処理部１７４は、上記特徴語抽出処理部１７１で抽出された各特徴語

について、種文書１６０１内での出現頻度を計数し、特徴語とその出現頻度を特徴ベクトル１６０５として出力する。以上が、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順である。

【００９８】

以上説明した様に本実施形態によれば、複合特徴語だけでなく、複合特徴語を構成する構成特徴語を特徴ベクトルの要素として類似度算出に使用する。この結果として、最長一致の様に“地図閲覧ソフト”を含むノイズの少ない類似文書検索を行なうと共に検索漏れの無い高精度な類似文書検索を実現することができる。

【００９９】

なお本実施形態では、登録対象文書や種文書を文書としたが、文章或いは文字列であっても構わない。また、本実施形態における特徴ベクトル生成処理では、処理対象中に出現する複合特徴語から複合特徴語及び複合特徴語に含まれる構成特徴語を全て抽出するものとして説明したが、全ての構成特徴語を抽出するのではなく一部を抽出するものとしても構わない。この場合、抽出する構成特徴語の指定方法としては、従来技術１の参照文献等に記載されているＩＤＦ(Inverted Document Frequency) が予め定められた閾値を越えるものだけを抽出するものとしても良いし、複合特徴語の中で先頭或いは末尾等の予め定められた位置を構成する特徴語だけを抽出するものとしても良い。

【０１００】

また本実施形態では、登録対象文書に対する特徴ベクトルを予め作成しておくものとしたが、文書の登録時には全文検索用インデクスを作成しておき、検索時に該当する全文検索用インデクスを参照することにより各登録対象文書に出現頻度を求め、類似度の算出を行なうものとしても良い。更に本実施形態では、特徴語の抽出に特徴語辞書を参照するものとして説明したが、辞書を用いずに特徴語を抽出する技術等を用いることも可能である。

【０１０１】

また、本実施形態では日本語における類似文書検索システムの例を説明したが、日本語に限らず他言語であっても構わない。すなわち、前述の日本語における

類似文書検索システムの場合には、種文書中に存在する複合特徴語及び該複合特徴語を構成する単語を類似度算出に使用することで検索漏れの無い類似文書検索を実現していたが、例えば英語等の様に単語の境界が明確な言語の場合には、複数の単語の組（一般にフレーズや熟語と呼ばれる）を複合特徴語として取り扱い、フレーズや熟語を用いた検索を行なう際に、本実施形態を適用することが可能となる。

【０１０２】

これにより、他言語においても意味のつながりのある単語の組の内容を考慮した類似度算出を行なうことができる様になり、検索漏れの少ない多言語対応の類似文書検索を提供することができる様になる。

【０１０３】

まず、従来技術１を英文対応類似文書検索システムに適用した場合の問題点について図１１を用いて説明する。

【０１０４】

図１１は従来技術１を英文対応類似文書検索システムに適用した場合の問題点を示す図である。本図では、文書５「This juice is made of carrot...」及び文書６「-- Carrot Juice -- 1. Cut carrot into some pieces...」が登録された文書データベースに対して、種文書「How to make carrot juice 」が入力された場合の例を表している。

【０１０５】

まず文書の登録処理として、ステップ２１０が実行され、各文書に対応する登録文書特徴ベクトル１７０２及び１７０３が生成される。本図に示した例では、文書５に対応する登録文書特徴ベクトル１７０２として“文書５（"carrot"、1）（"juice"、1）”が生成され、文書６に対応する登録文書特徴ベクトル１７０３として“文書６（"carrot juice"、1）（"carrot"、1）”が生成される。

【０１０６】

次に類似文書の検索処理として、ステップ２２０が実行され、種文書に対応する種文書特徴ベクトル１７０６が生成される。本図に示した例では、種文書特徴ベクトル１７０６として、“種文書（"carrot juice"、1）”が生成される。

【０１０７】

そしてステップ２２２において、種文書に対する各登録文書の類似度を算出する。この結果、類似度算出結果１７０７が出力される。本図に示した例では、数８及び数９に示す様に、文書５の類似度０．０００及び文書６の類似度０．７１０と算出される。

【０１０８】

【数８】

数8

$$\frac{1\times 0}{\sqrt{1^2}\times\sqrt{1^2+1^2}} = \frac{0}{\sqrt{2}} = 0.000$$

【０１０９】

【数９】

数9

$$\frac{1\times 1}{\sqrt{1^2}\times\sqrt{1^2+1^2}} = \frac{1}{\sqrt{2}} = 0.710$$

【０１１０】

この結果、文書５の内容は種文書に対して関連があるにも関わらず、従来技術１では文書５の内容は種文書に対して全く類似していないものと算出されてしまう。

【０１１１】

これは、種文書の特徴ベクトルの要素として抽出される特徴語が複数の単語の組で構成されているにもかかわらず、該特徴語のみを特徴ベクトルの要素として類似度算出に用いた為に、特徴語を構成する各単語の持つ個々の概念が類似度に反映されないことによるものである。

【０１１２】

すなわち、ノイズ等を減らす為に複数の単語の組である"carrot juice"等を特徴語とした場合には、"carrot juice"を含む文書６に対する検索精度が向上し、"carrot juice"を含まない登録文書は検索されなくなるが、その特徴語を構成する各単語それぞれを含む登録文書の文書５に対して類似度が付与されず、検索漏れが発生してしまうことになる。

【０１１３】

以上説明した様に従来技術１を英文対応類似文書検索システムに適用した場合にも、日本語の場合と同様の問題が生じてしまうことになる。上記問題に対し、本実施形態を英文対応類似文書検索システムに適用することにより、日本語の場合と同様に解決することができる様になる。

【０１１４】

以下、図１２に本実施形態を適用した英文対応類似文書検索システムの処理概要を示す。

【０１１５】

図１２は本実施形態の英文対応類似文書検索システムの処理概要を示す図である。図１２は、文書５「This juice is made of carrot...」及び文書６「-- Carrot juice -- 1. Cut carrot into some pieces...」が登録された文書データベースに対して、種文書「How to make carrot juice」が入力された場合の例を表している。

【０１１６】

まず文書の登録処理のステップ２１０では、各文書に対応する登録文書特徴ベクトル１７０２及び１７０３を生成する。本図に示した例では、文書５に対応する登録文書特徴ベクトル１７０２として“文書５（"carrot"、1）（"juice"、1）”を生成し、文書６に対応する登録文書特徴ベクトル１７０３ａとして“文書６（"carrot juice"、1）（"carrot"、2）（"juice"、1）”を生成する。

【０１１７】

次に文書の検索処理のステップ２２０では、種文書に対応する種文書特徴ベクトル１７０６ａを生成する。本図に示した例では、種文書特徴ベクトル１７０６ａとして、“種文書（"carrot juice"、1）（"carrot"、1）（"juice"、1）”を生

成する。

【０１１８】

そしてステップ２２２において、種文書に対する各登録文書の類似度を算出する。この結果、類似度算出結果１７０７ａを出力する。数１０及び数１１に示す様に、本図に示した例では文書５の類似度０．８１６及び文書６の類似度０．９４３と算出される。

【０１１９】

【数１０】

数10

$$\frac{1\times0+1\times1+1\times1}{\sqrt{1^2+1^2+1^2}\times\sqrt{1^2+1^2}}$$

$$=\frac{2}{\sqrt{3}\times\sqrt{2}}=0.816$$

【０１２０】

【数１１】

数11

$$\frac{1\times1+1\times2+1\times1}{\sqrt{1^2+1^2+1^2}\times\sqrt{1^2+2^2+1^2}}$$

$$=\frac{4}{\sqrt{3}\times\sqrt{6}}=0.943$$

【０１２１】

以上が本実施形態を適用した英文対応類似文書検索システムの処理概要である。前記の様に本実施形態を適用した英文対応類似文書検索システムにおいても複合特徴語を考慮することにより、従来技術１では検索することができない文書５を検索することができる様になる。

【０１２２】

以上説明した様に本実施形態の類似文書検索システムによれば、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するので、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を

精度良く検索することが可能である。

【０１２３】

（実施形態２）

以下に複合特徴語から抽出された構成特徴語の登録文書内での出現距離を考慮した重み付けを行なう実施形態２の類似文書検索システムについて説明する。

【０１２４】

本実施形態を適用した類似文書検索システムでは、複合特徴語から抽出された構成特徴語の登録文書内での出現距離を考慮した重み付けを行なうものであり、種文書の同一複合特徴語から抽出された関連性の高い構成特徴語が、関連の高い出現関係にある登録文書に対して高い類似度を付与することにより、より内容の近い登録文書を検索し、高精度な検索結果を得ることができる様になる。

【０１２５】

図１３は本実施形態の特徴ベクトル生成処理部１７０ａの構成を示す図である。図１３に示す様に本実施形態の類似文書検索システムは出現位置取得処理部１９００を有している。出現位置取得処理部１９００は、特徴語抽出処理部１７１で抽出された各特徴語について、処理対象文書内での出現位置を取得する処理部である。

【０１２６】

類似文書検索システムを出現位置取得処理部１９００として機能させる為のプログラムは、ＣＤ－ＲＯＭ等の記録媒体に記録され磁気ディスク等に格納された後、メモリにロードされて実行されるものとする。なお前記プログラムを記録する記録媒体はＣＤ－ＲＯＭ以外の他の記録媒体でも良い。また前記プログラムを当該記録媒体から情報処理装置にインストールして使用しても良いし、ネットワークを通じて当該記録媒体にアクセスして前記プログラムを使用するものとしても良い。

【０１２７】

図１４は本実施形態の種文書類似度算出処理部１３１ａの構成を示す図である。図１４に示す様に本実施形態の類似文書検索システムは重み係数算出処理部２０００を有している。重み係数算出処理部２０００は、種文書から抽出された構

成特徴語に一致する登録文書の特徴語について、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を算出する処理部である。

【０１２８】

類似文書検索システムを重み係数算出処理部２０００として機能させる為のプログラムは、ＣＤ－ＲＯＭ等の記録媒体に記録され磁気ディスク等に格納された後、メモリにロードされて実行されるものとする。なお前記プログラムを記録する記録媒体はＣＤ－ＲＯＭ以外の他の記録媒体でも良い。また前記プログラムを当該記録媒体から情報処理装置にインストールして使用しても良いし、ネットワークを通じて当該記録媒体にアクセスして前記プログラムを使用するものとしても良い。

【０１２９】

本実施形態は、実施形態１（図１）とほぼ同様の構成を取るが、特徴ベクトル生成処理部１７０及び種文書類似度算出処理部１３１の構成が異なる。特徴ベクトル生成処理部１７０ａでは、図１３に示す様に、出現位置取得処理部１９００が用いられる。また、種文書類似度算出処理部１３１ａでは、図１４に示す様に重み係数算出処理部２０００が用いられる。

【０１３０】

以下、本実施形態における処理手順の内、まず実施形態１とは異なる特徴ベクトル生成処理部１７０ａの処理手順について、図１５に示すＰＡＤ図を用いて説明する。

【０１３１】

図１５は本実施形態の特徴ベクトル生成処理部１７０ａの処理内容を示す図である。ここで、実施形態１における特徴ベクトル生成処理部１７０（図４）と異なる点は、出現位置取得ステップ２１００が加わるだけである。他の処理ステップの処理手順は、実施形態１で説明した通りである。

【０１３２】

出現位置取得ステップ２１００では、出現位置取得処理部１９００を起動し、ワークエリア１４１に格納された各単語の、処理対象文書における出現位置を取得する。以上が、特徴ベクトル生成処理部１７０ａの処理手順である。

【０１３３】

次に、本実施形態における種文書類似度算出処理部１３１ａの処理手順について、図１６に示すＰＡＤ図を用いて説明する。

【０１３４】

図１６は本実施形態の種文書類似度算出処理部１３１ａの処理内容を示す図である。ここで、実施形態１における種文書類似度算出処理部１３１（図７）と異なる点は、重み係数算出ステップ２２００が加わるだけである。他の処理ステップの処理手順は、実施形態１で説明した通りである。

【０１３５】

重み係数算出ステップ２２００では、重み係数算出処理部２０００を起動し、種文書特徴ベクトルの各要素の内、同一の複合特徴語から抽出された構成特徴語の組に対して重み係数を算出し、種文書特徴ベクトルの要素に乗じる。以上が、種文書類似度算出処理部１３１ａの処理手順である。

【０１３６】

以下、本実施形態における類似文書検索システムの具体的な処理手順を図１７～図１９を用いて説明する。まず、本実施形態における類似文書検索システムにおける文書の登録処理について、図１７を用いて説明する。

【０１３７】

図１７は本実施形態の文書登録処理の概要を示す図である。図１７では、文書３「地図情報閲覧ソフトを開発、発売したＡ社は、・・・」及び文書４「多くの地図閲覧ソフトが発売されているが、・・・」が文書データベースに登録される場合の処理の流れを表している。

【０１３８】

まず登録文書取得処理部１２０は、文書３及び文書４を読み込み、ワークエリア１４１に格納する。次に特徴ベクトル生成処理部１７０ａは、登録対象の文書３及び文書４に対して対応する登録文書特徴ベクトル２３００及び２３０１を作成し、ワークエリア１４１に格納する。

【０１３９】

本図に示した例では、文書３に対応する登録文書特徴ベクトル２３００として

“文書３（”地図”、1）[1]、（”閲覧”、1）[5]、（”ソフト”、1）[7]、（”発売”、1）[14]”が生成され、文書４に対応する登録文書特徴ベクトル２３０１として“文書４（”地図閲覧ソフト”、1）[4]、（”地図”、1）[4]、（”閲覧”、1）[6]、（”ソフト”、1）[8]、（”発売”、1）[12]”が生成される。なお、ここで“（”地図”、1）[1]”の丸括弧（）内は特徴語“地図”が１回出現することを表し、角括弧[]内の“１”は特徴語“地図”の文字位置が１であることを表している。

【０１４０】

そして、登録文書特徴ベクトル登録処理部１２１は、ワークエリア１４１上の登録文書特徴ベクトルを登録文書特徴ベクトルファイル１８０として格納する。以上が、本実施形態に示した類似文書検索システムにおける文書の登録処理である。

【０１４１】

次に、本実施形態における類似文書検索システムにおける類似文書の検索処理について、図１８を用いて説明する。

【０１４２】

図１８は本実施形態の類似文書の検索処理の処理内容を示す図である。図１８では、種文書「最新の地図閲覧ソフトについて」が入力された場合の例を表している。まず、検索条件解析処理部１３０は、検索条件で指定された種文書を取得し、ワークエリア１４１に格納する。

【０１４３】

そして、特徴ベクトル生成処理部１７０ａは、ワークエリア１４１に格納された種文書に対応する種文書特徴ベクトル２４００を生成し、ワークエリア１４１に格納する。

【０１４４】

次に、登録文書特徴ベクトル読込処理部１６０は、前記文書の登録処理で作成された登録文書特徴ベクトルファイル１８０を読み込み、登録文書特徴ベクトル２３００及び２３０１をワークエリア１４１に格納する。

【０１４５】

そして、重み係数算出処理部２０００は、種文書特徴ベクトル２４００の各要

素が構成特徴語であるかを判定し、該要素がある複合特徴語の構成特徴語である場合には数１２に基づいて重みを算出し、重み係数２４０１として出力する。

【０１４６】

【数１２】

数12

種文書特徴ベクトルの構成特徴語Aに対する重み係数

$$= 1 - \frac{\min(\text{同親構成特徴語との最短距離}, C)}{C}$$

【０１４７】

本図では、数１２（定数Ｃ＝２０とした）に基づき各要素の重みを算出した場合の例を表している。すなわち、種文書特徴ベクトル２４００の要素“地図”は複合特徴語“地図閲覧ソフト”の構成特徴語であるから、同じ複合特徴語“地図閲覧ソフト”から抽出された他の構成特徴語（以下、同親構成特徴語と呼ぶ）である“閲覧”、“ソフト”との最短距離を取得する。本図に示した例では文書３において、“地図”に対する“閲覧”、“ソフト”の最短距離は“４”であるから、重み係数“０．８０”が算出されている。

【０１４８】

そして、類似度算出処理部１６１において、前記ステップ１７０ａで生成された種文書特徴ベクトル２４００と登録文書特徴ベクトル２３００及び２３０１のなす角度の余弦が数１３及び数１４の様に算出され、種文書に対する登録文書の類似度算出結果２４０２が出力される。

【０１４９】

【数１３】

数13

$$\frac{1\times1\times0 + 0.8\times1\times1 + 0.9\times1\times1 + 0.9\times1\times1}{\sqrt{(1\times1)^2 + (0.8\times1)^2 + (0.9\times1)^2 + (0.9\times1)^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2}}$$

$$= \frac{2.6}{\sqrt{3.26}\sqrt{4}} = 0.720$$

【０１５０】

【数１４】

数14

$$\frac{1\times1\times1+0.9\times1\times1+0.9\times1\times1+0.9\times1\times1}{\sqrt{(1\times1)^2+(0.9\times1)^2+(0.9\times1)^2+(0.9\times1)^2}\times\sqrt{1^2+1^2+1^2+1^2+1^2}}$$

$$=\frac{3.7}{\sqrt{3.43}\sqrt{5}}=0.893$$

【０１５１】

以上が、本実施形態における類似文書検索システムにおける類似文書の検索処理手順である。

【０１５２】

次に、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順について図１９を用いて説明する。

【０１５３】

図１９は本実施形態の特徴ベクトルの生成処理の処理内容を示す図である。図１９では、種文書「最新の地図閲覧ソフトについて」が入力された場合の例に、検索特徴ベクトルが作成される手順を表している。

【０１５４】

まず、文書解析処理部１７２は、ワークエリア１４１に格納された処理対象文書である種文書１６０１“最新の地図閲覧ソフトについて”から特徴語候補１６０２“地図閲覧ソフト”を抽出する。

【０１５５】

そして、複合特徴語判定処理部１７３は、特徴語候補１６０２“地図閲覧ソフト”が複数の特徴語で構成される特徴語かを判定する。この結果、特徴語候補１６０２“地図閲覧ソフト”は複数の特徴語“地図”、“閲覧”、“ソフト”から構成されるものと判定され、複合特徴語と判定される。

【０１５６】

次に、特徴語抽出処理部１７１では、上記複合特徴語判定処理部１７３の結果、複合特徴語と判定された“地図閲覧ソフト”から、これを構成する特徴語１６０４“地図”、“閲覧”、“ソフト”を抽出する。そして、出現頻度計数処理部

１７４は、上記特徴語抽出処理で抽出された各特徴語について、種文書１６０１内での出現頻度を計数する。

【０１５７】

そして、出現位置取得処理部１９００は、上記特徴語抽出処理部１７１で抽出された各特徴語について、種文書１６０１内での出現位置を取得し、特徴ベクトル２５００として出力する。以上が、本実施形態における類似文書検索システムにおける特徴ベクトルの生成処理手順である。

【０１５８】

以上説明した様に、本実施形態によれば、種文書から抽出された複合特徴語の構成特徴語間の距離を考慮することにより、登録文書内での単語間の関係を考慮した高精度な類似度算出を行なうことができる。すなわち、複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索することにより、検索漏れの無い高精度な類似文書検索が可能となるが、その際に構成特徴語間の距離を考慮して重み付けを行なことにより、種文書との関連が低い登録文書の類似度を下げて検索時のノイズを削減することが可能である。

【０１５９】

なお、本実施形態における特徴ベクトル生成処理部１７０ａでは、複合特徴語及び複合特徴語から抽出された構成特徴語の両方を特徴語として抽出していたが、構成特徴語だけを抽出するものとして良い。この場合、重み係数算出や類似度算出に使用される特徴語の要素数が削減される為、より高速な検索を実現することができる。

【０１６０】

また、本実施形態における特徴ベクトル生成処理部１７０ａでは、各特徴語の出現位置取得処理部１９００を出現頻度計数処理部１７４の後に実施するものとしたが、種文書解析処理部１７２の実施時に各特徴語候補を抽出するのに合わせて、各特徴語候補文字列の出現位置を抽出しておくものとしても良い。

【０１６１】

更に、本実施形態における特徴ベクトル２５００では、各要素に対応して出現頻度及び出現位置を格納するものとしたが、種文書に対する特徴ベクトル作成処

理では同親構成特徴語をまとめて一つの要素として管理するものとしても良い。この様にすることにより、重み係数算出処理時に各要素が構成特徴語か否かを判断する必要がない為、より高速な検索を実現することができる。

【０１６２】

以上説明した様に本実施形態の類似文書検索システムによれば、同一の複合特徴語から抽出された他の構成特徴語との間の距離に応じた重み係数を乗じた類似度を算出するので、検索漏れが無くノイズの少ない高精度な類似文書検索を実現することが可能である。

【０１６３】

【発明の効果】

本発明によれば複合特徴語及びその複合特徴語を構成する構成特徴語を含む文書を類似文書として検索するので、検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能である。

【図面の簡単な説明】

【図１】

実施形態１の類似文書検索システムの概略構成を示す図である。

【図２】

実施形態１のシステム制御処理部１１０の処理内容を示す図である。

【図３】

実施形態１の登録制御処理部１１１の処理内容を示す図である。

【図４】

実施形態１の特徴ベクトル生成処理部１７０の処理内容を示す図である。

【図５】

実施形態１の特徴語抽出処理部１７１の処理内容を示す図である。

【図６】

実施形態１の検索制御処理部１１２の処理内容を示す図である。

【図７】

実施形態１の種文書類似度算出処理部１３１の処理内容を示す図である。

【図８】

実施形態１の文書の登録処理の処理内容を示す図である。

【図９】

実施形態１の類似文書の検索処理の処理内容を示す図である。

【図１０】

実施形態１の特徴ベクトルの生成処理の処理内容を示す図である。

【図１１】

従来技術１を英文対応類似文書検索システムに適用した場合の問題点を示す図である。

【図１２】

実施形態１の英文対応類似文書検索システムの処理概要を示す図である。

【図１３】

実施形態２の特徴ベクトル生成処理部１７０ａの構成を示す図である。

【図１４】

実施形態２の種文書類似度算出処理部１３１ａの構成を示す図である。

【図１５】

実施形態２の特徴ベクトル生成処理部１７０ａの処理内容を示す図である。

【図１６】

実施形態２の種文書類似度算出処理部１３１ａの処理内容を示す図である。

【図１７】

実施形態２の文書登録処理の概要を示す図である。

【図１８】

実施形態２の類似文書の検索処理の処理内容を示す図である。

【図１９】

実施形態２の特徴ベクトルの生成処理の処理内容を示す図である。

【図２０】

従来技術１の処理手順の一例を示す図である。

【図２１】

従来技術１における特徴ベクトル生成処理の一例を示す図である。

【図２２】

従来技術１の概要を示す図である。

【図２３】

従来技術１の問題点を示す図である。

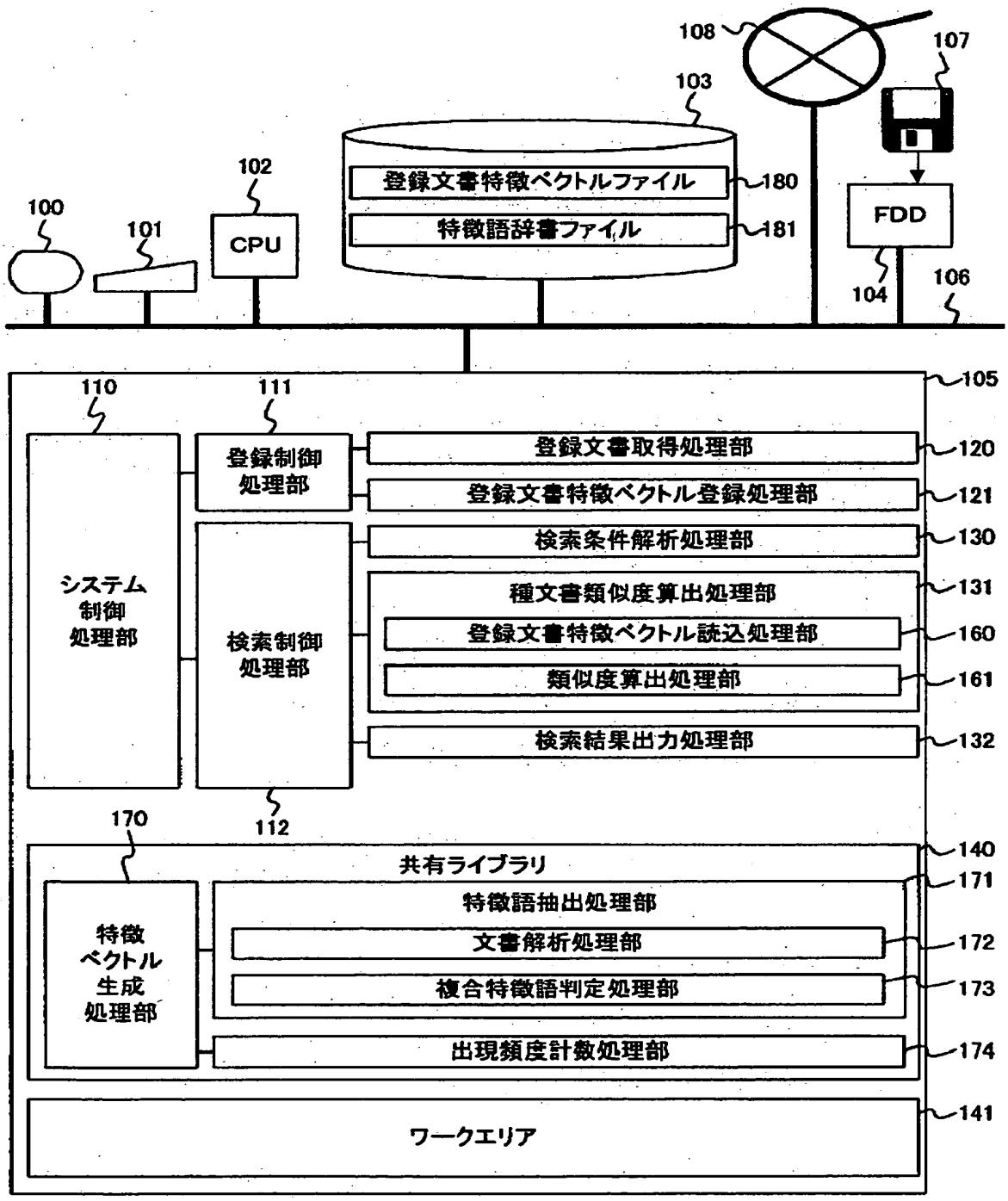
【符号の説明】

１００…ディスプレイ、１０１…キーボード、１０２…ＣＰＵ、１０３…磁気ディスク装置、１０４…ＦＤＤ、１０５…主メモリ、１０６…バス、１０７…フロッピディスク、１０８…ネットワーク、１４０…共有ライブラリ、１４１…ワークエリア、１８０…登録文書特徴ベクトルファイル、１８１…特徴語辞書ファイル、１１０…システム制御処理部、１１１…登録制御処理部、１１２…検索制御処理部、１２０…登録文書取得処理部、１２１…登録文書特徴ベクトル登録処理部、１３０…検索条件解析処理部、１３１…種文書類似度算出処理部、１３２…検索結果出力処理部、１６０…登録文書特徴ベクトル読込処理部、１６１…類似度算出処理部、１７０…特徴ベクトル生成処理部、１７１…特徴語抽出処理部、１７２…文書解析処理部、１７３…複合特徴語判定処理部、１７４…出現頻度計数処理部、１６０１…種文書、１６０２…特徴語候補、１６０２及び１６０３…登録文書特徴ベクトル、１６０３…特徴ベクトル、１６０４…特徴語、１６０５…特徴ベクトル、１７００及び１７０１…登録文書、１７０２及び１７０３…登録文書特徴ベクトル、１７０５…種文書、１７０６…種文書特徴ベクトル、１７０７…類似度算出結果、１９００…出現位置取得処理部、２０００…重み係数算出処理部、２３００及び２３０１…登録文書特徴ベクトル、２４００…種文書特徴ベクトル、２４０１…重み係数、２４０２…類似度算出結果、２５００…特徴ベクトル、４０１及び４０２…登録文書、４０３及び４０４…登録文書特徴ベクトル、４０５…特徴語辞書、４０６…種文書、４０７…種文書特徴ベクトル、４０８…類似度算出結果。

【書類名】　　図面

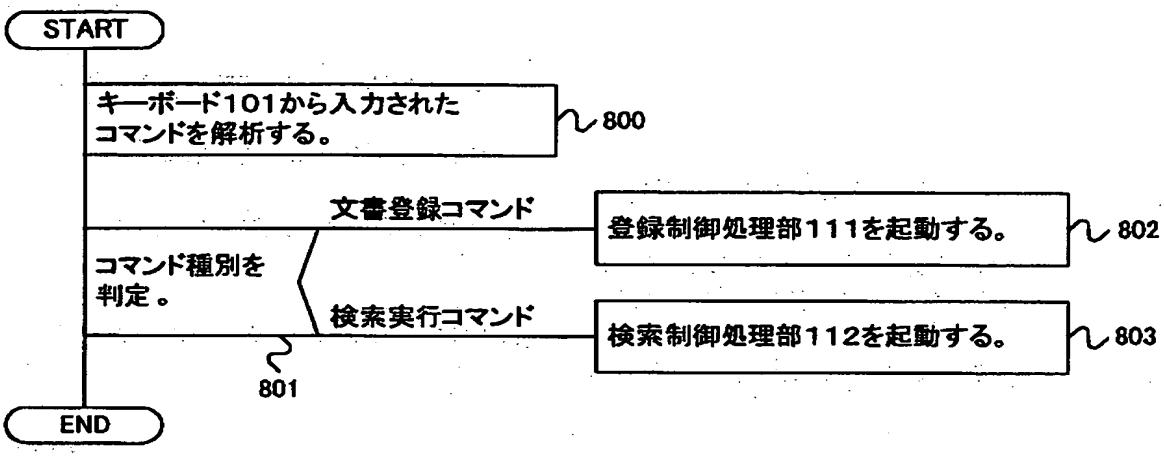
【図１】

図1


108
107
103
102
登録文書特徴ベクトルファイル
180
特徴語辞書ファイル
181
100
101
CPU
FDD
104
106
105
110
111
システム制御処理部
登録制御処理部
登録文書取得処理部
120
登録文書特徴ベクトル登録処理部
121
検索制御処理部
検索条件解析処理部
130
種文書類似度算出処理部
131
登録文書特徴ベクトル読込処理部
160
類似度算出処理部
161
検索結果出力処理部
132
112
170
共有ライブラリ
140
特徴ベクトル生成処理部
特徴語抽出処理部
171
文書解析処理部
172
複合特徴語判定処理部
173
出現頻度計数処理部
174
ワークエリア
141

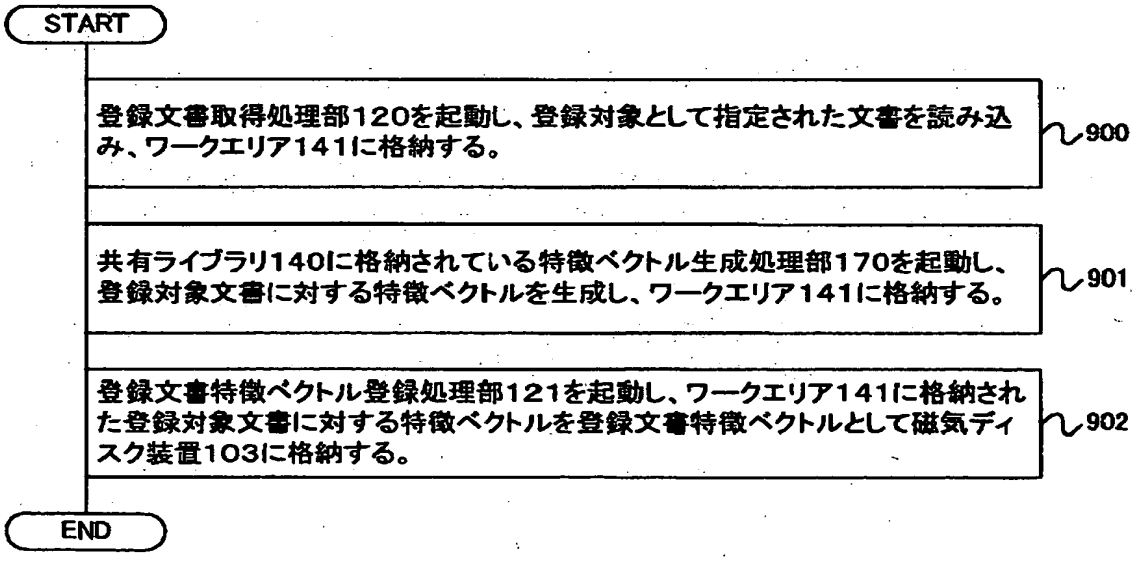
【図２】

図2


START
キーボード101から入力された
コマンドを解析する。
800
文書登録コマンド
登録制御処理部111を起動する。
802
コマンド種別を
判定。
検索実行コマンド
検索制御処理部112を起動する。
803
801
END

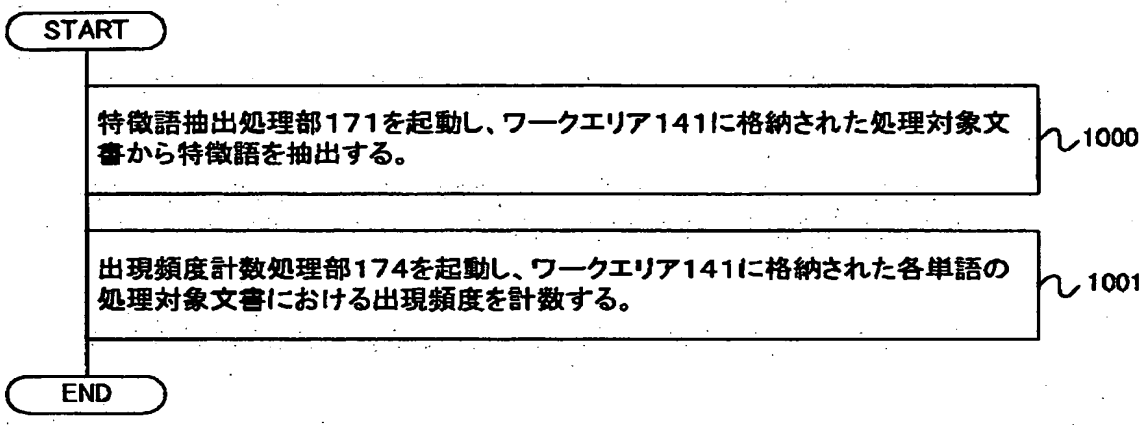
【図３】

図3


START
登録文書取得処理部120を起動し、登録対象として指定された文書を読み込み、ワークエリア141に格納する。
900
共有ライブラリ140に格納されている特徴ベクトル生成処理部170を起動し、登録対象文書に対する特徴ベクトルを生成し、ワークエリア141に格納する。
901
登録文書特徴ベクトル登録処理部121を起動し、ワークエリア141に格納された登録対象文書に対する特徴ベクトルを登録文書特徴ベクトルとして磁気ディスク装置103に格納する。
902
END

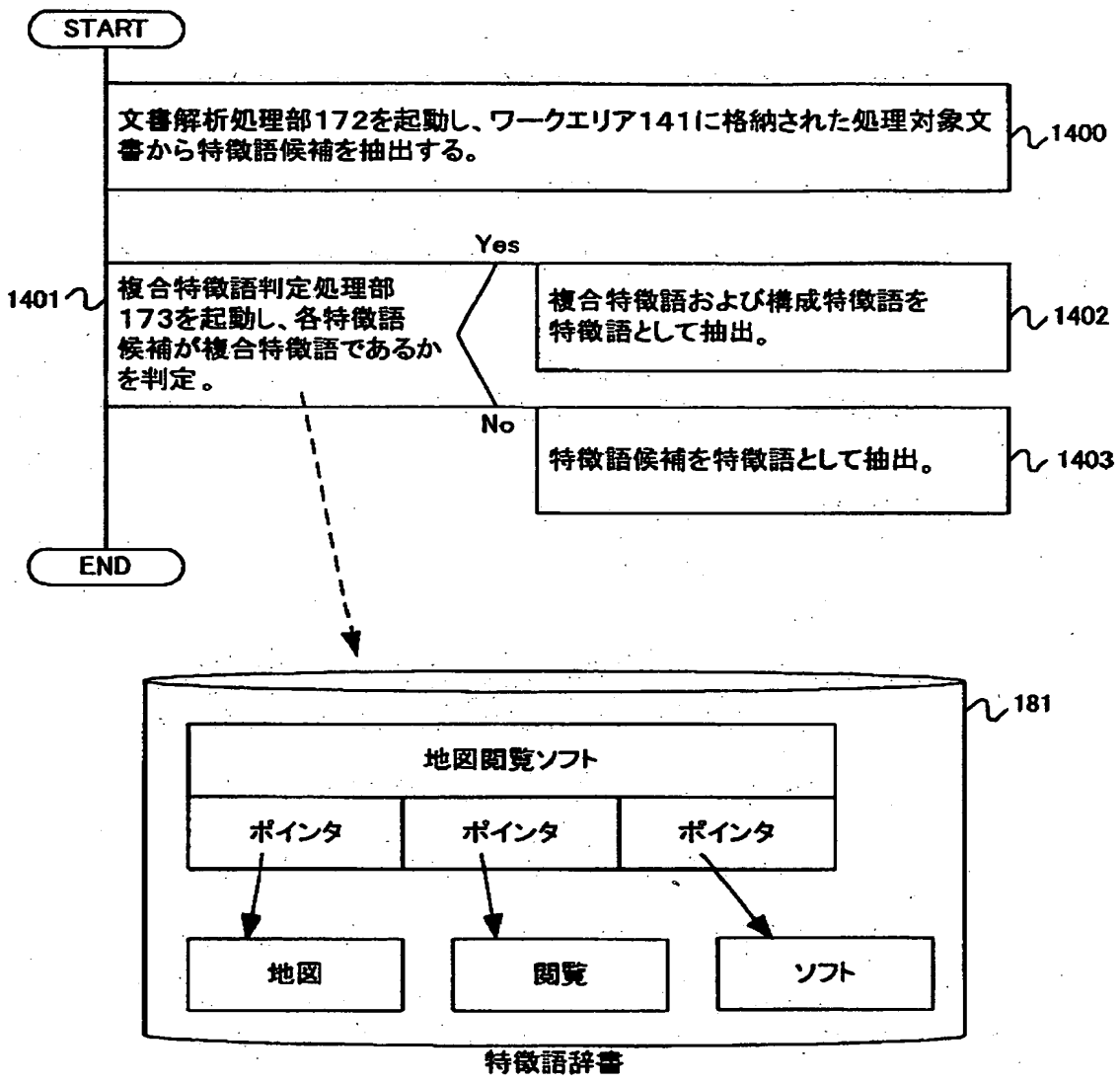
【図４】

図4


START
特徴語抽出処理部171を起動し、ワークエリア141に格納された処理対象文書から特徴語を抽出する。
1000
出現頻度計数処理部174を起動し、ワークエリア141に格納された各単語の処理対象文書における出現頻度を計数する。
1001
END

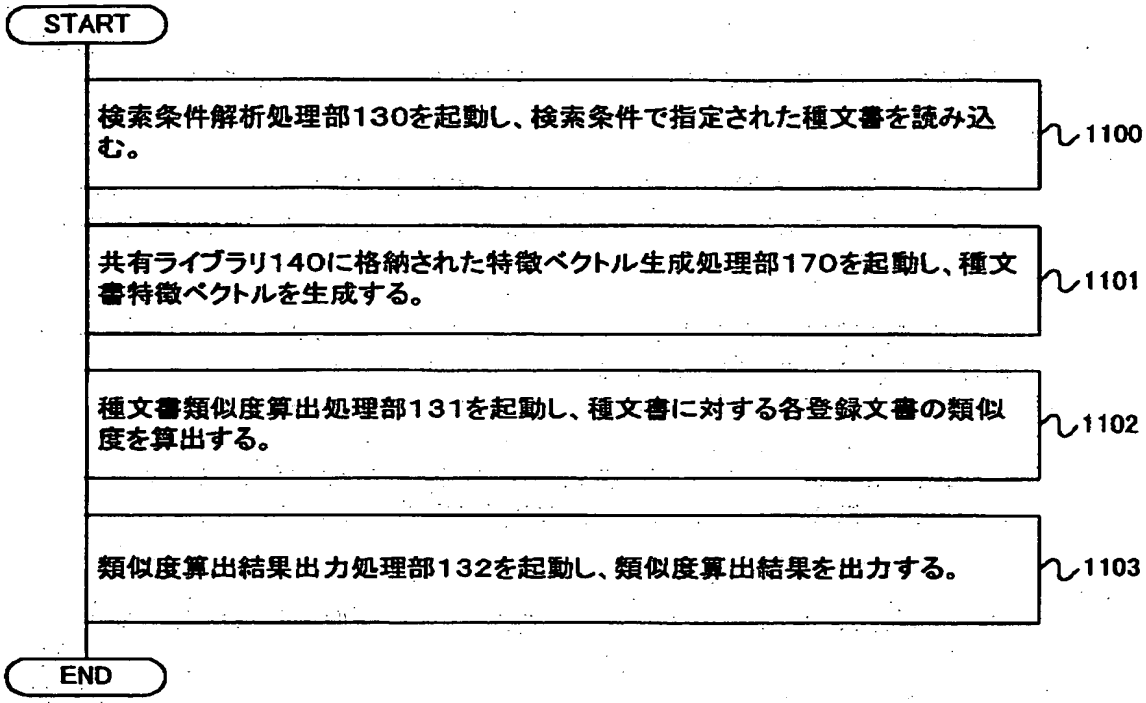
【図５】

図5


START
文書解析処理部172を起動し、ワークエリア141に格納された処理対象文書から特徴語候補を抽出する。
1400
Yes
1401
複合特徴語判定処理部173を起動し、各特徴語候補が複合特徴語であるかを判定。
複合特徴語および構成特徴語を特徴語として抽出。
1402
No
特徴語候補を特徴語として抽出。
1403
END
181
地図閲覧ソフト
ポインタ
ポインタ
ポインタ
地図
閲覧
ソフト
特徴語辞書

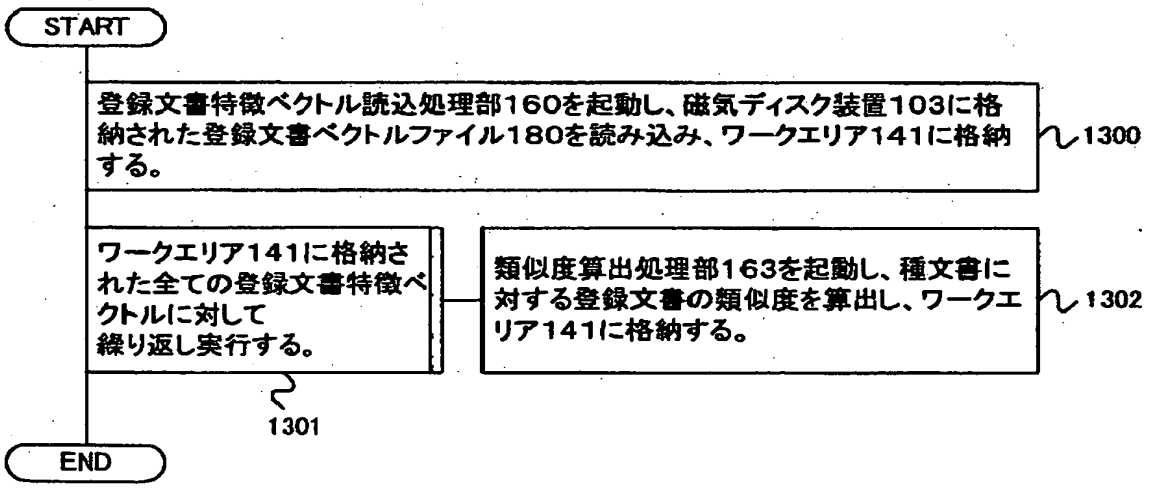
【図６】

図6


START
検索条件解析処理部130を起動し、検索条件で指定された種文書を読み込む。
1100
共有ライブラリ140に格納された特徴ベクトル生成処理部170を起動し、種文書特徴ベクトルを生成する。
1101
種文書類似度算出処理部131を起動し、種文書に対する各登録文書の類似度を算出する。
1102
類似度算出結果出力処理部132を起動し、類似度算出結果を出力する。
1103
END

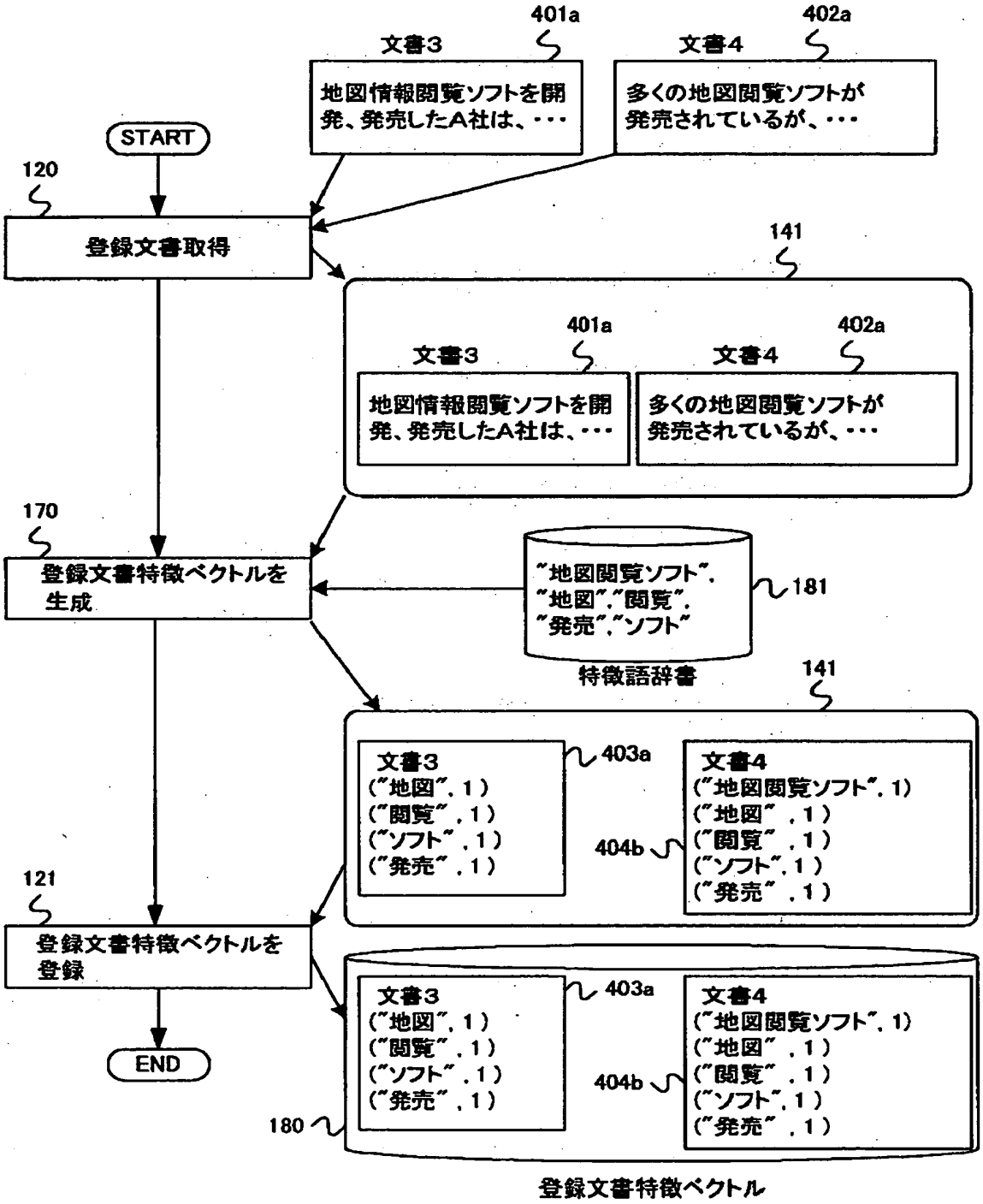
【図７】

図7


START
登録文書特徴ベクトル読込処理部160を起動し、磁気ディスク装置103に格納された登録文書ベクトルファイル180を読み込み、ワークエリア141に格納する。
1300
ワークエリア141に格納された全ての登録文書特徴ベクトルに対して
繰り返し実行する。
1301
類似度算出処理部163を起動し、種文書に対する登録文書の類似度を算出し、ワークエリア141に格納する。
1302
END

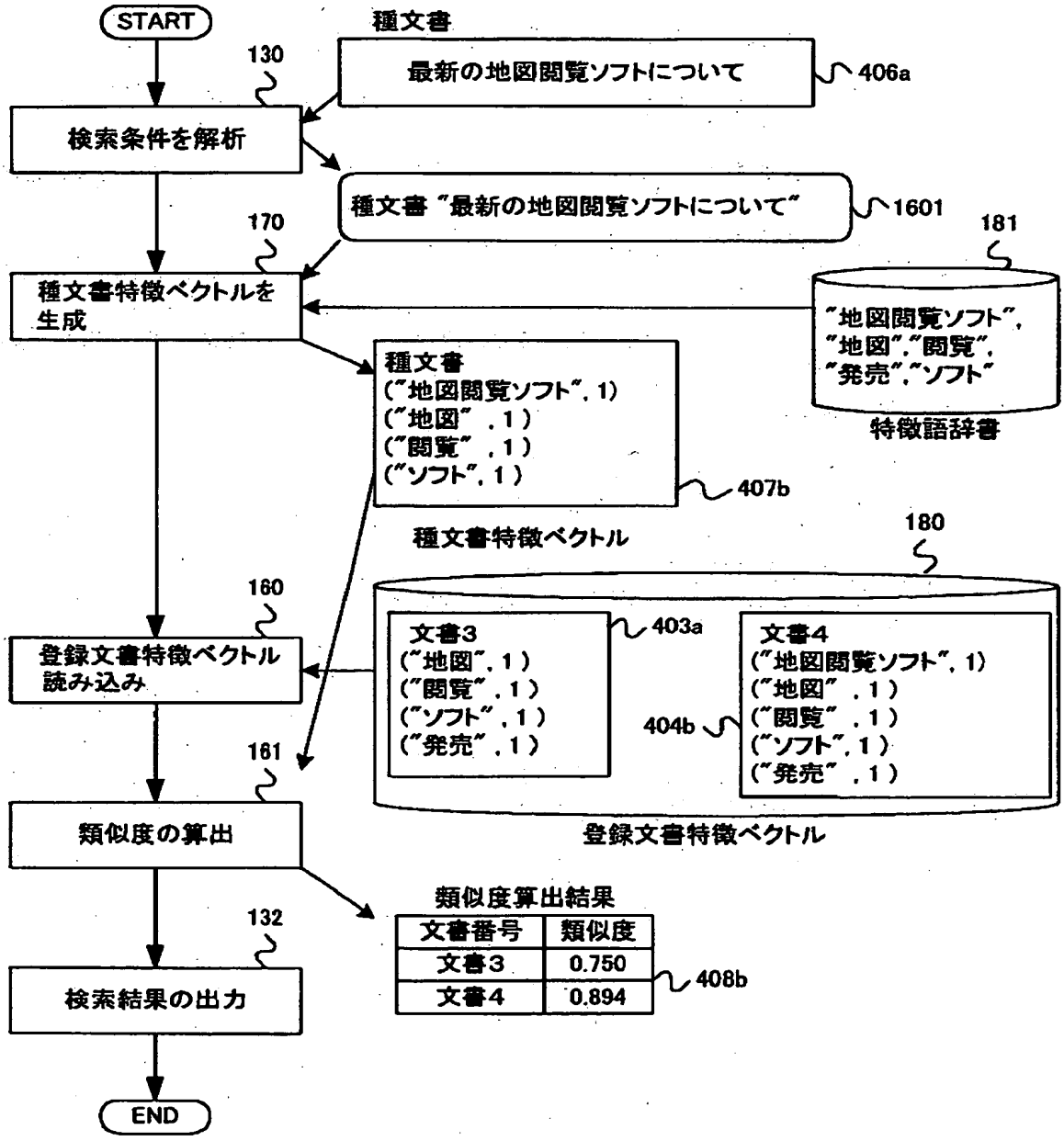
【図８】

図8


401a
402a
文書3
文書4
地図情報閲覧ソフトを開発、発売したA社は、・・・
多くの地図閲覧ソフトが発売されているが、・・・
START
120
登録文書取得
141
401a
402a
文書3
文書4
地図情報閲覧ソフトを開発、発売したA社は、・・・
多くの地図閲覧ソフトが発売されているが、・・・
170
登録文書特徴ベクトルを生成
"地図閲覧ソフト","地図","閲覧","発売","ソフト"
181
特徴語辞書
141
文書3
("地図",1)
("閲覧",1)
("ソフト",1)
("発売",1)
403a
404b
文書4
("地図閲覧ソフト",1)
("地図",1)
("閲覧",1)
("ソフト",1)
("発売",1)
121
登録文書特徴ベクトルを登録
END
180
登録文書特徴ベクトル

【図９】

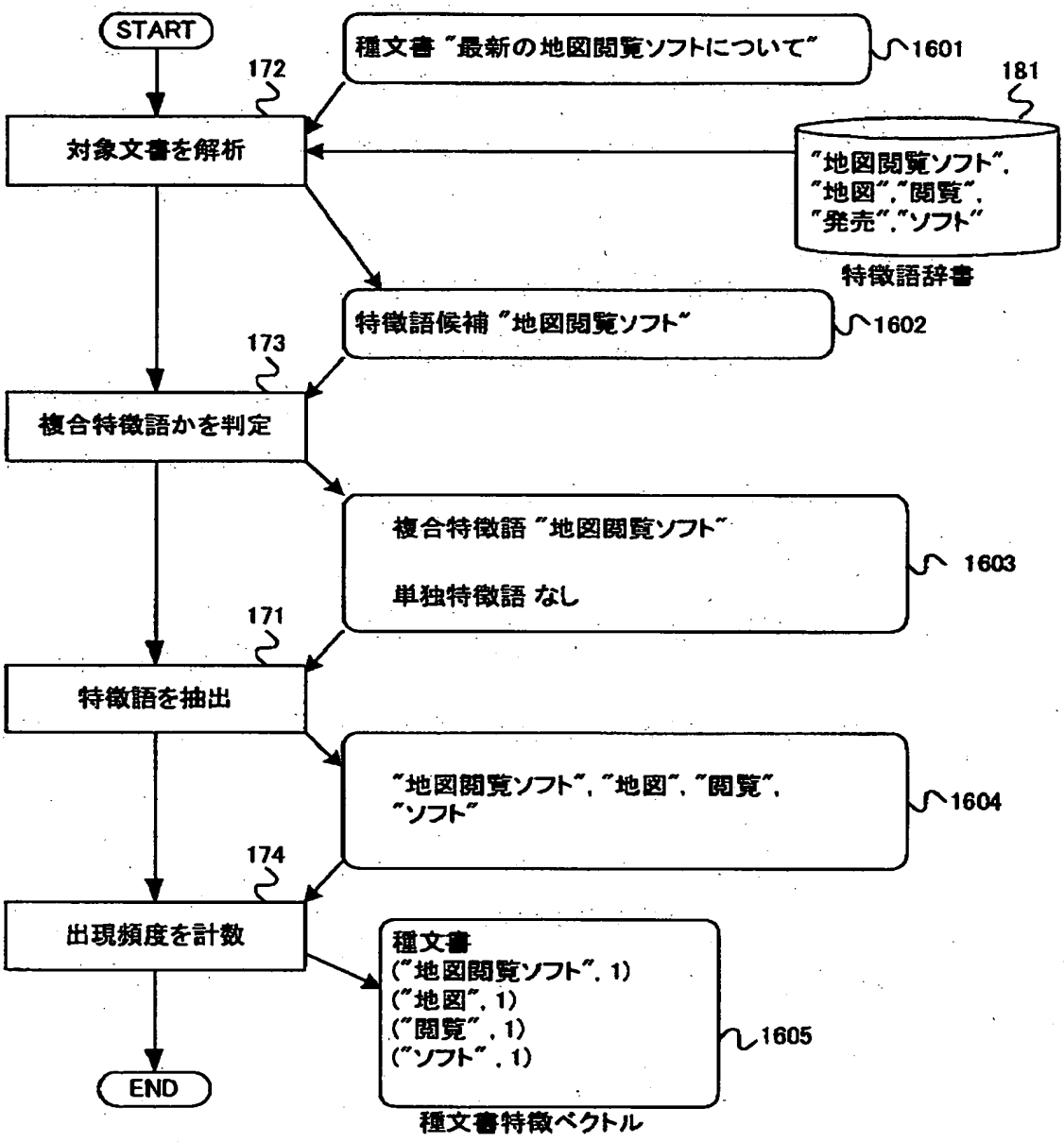
図9



文書番号	類似度
文書3	0.750
文書4	0.894

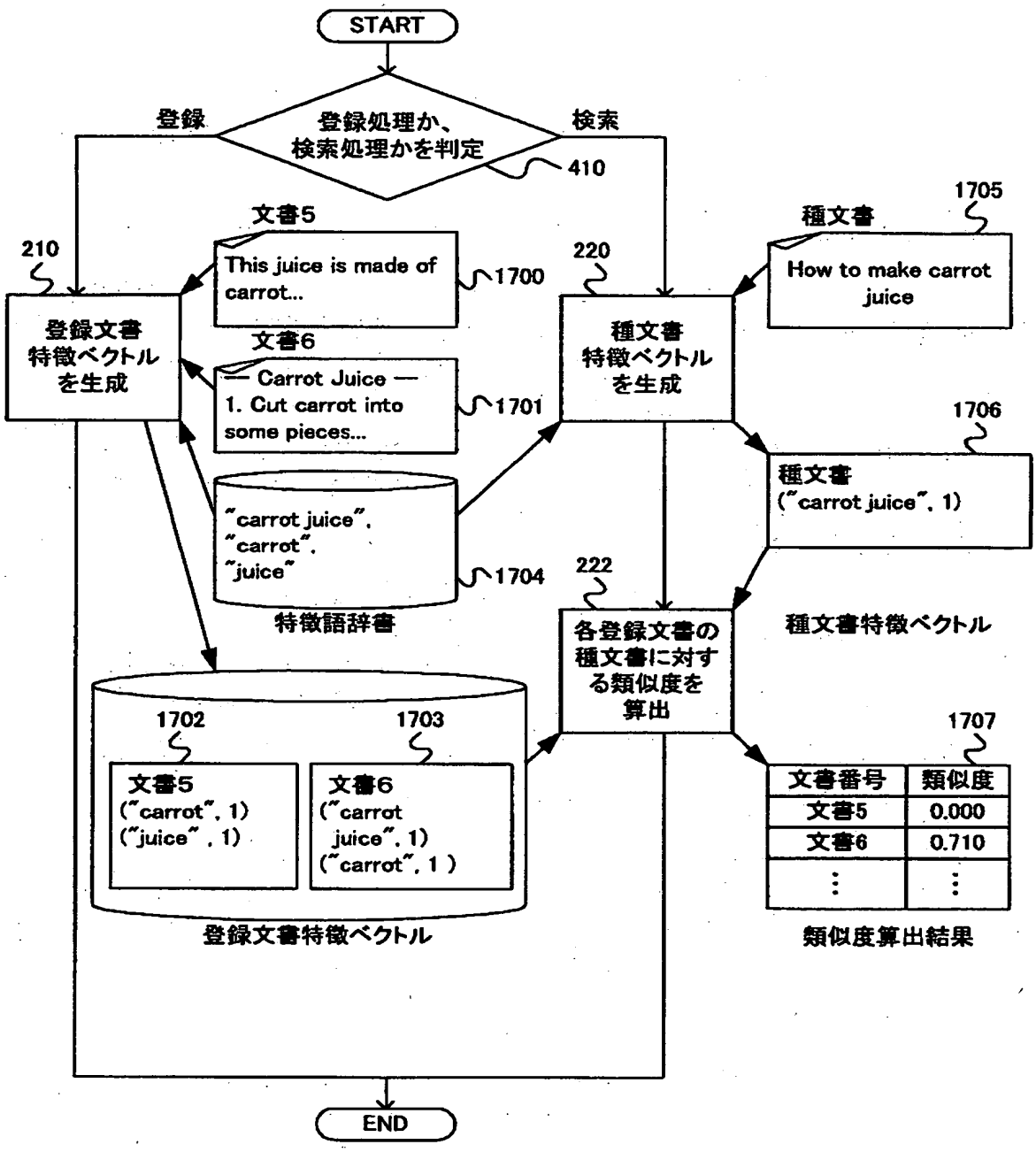
【図１０】

図10


START
種文書 "最新の地図閲覧ソフトについて"
1601
172
181
対象文書を解析
"地図閲覧ソフト", "地図", "閲覧", "発売", "ソフト"
特徴語辞書
特徴語候補 "地図閲覧ソフト"
1602
173
複合特徴語かを判定
複合特徴語 "地図閲覧ソフト"
単独特徴語 なし
1603
171
特徴語を抽出
"地図閲覧ソフト", "地図", "閲覧", "ソフト"
1604
174
出現頻度を計数
種文書
("地図閲覧ソフト", 1)
("地図", 1)
("閲覧", 1)
("ソフト", 1)
1605
種文書特徴ベクトル
END

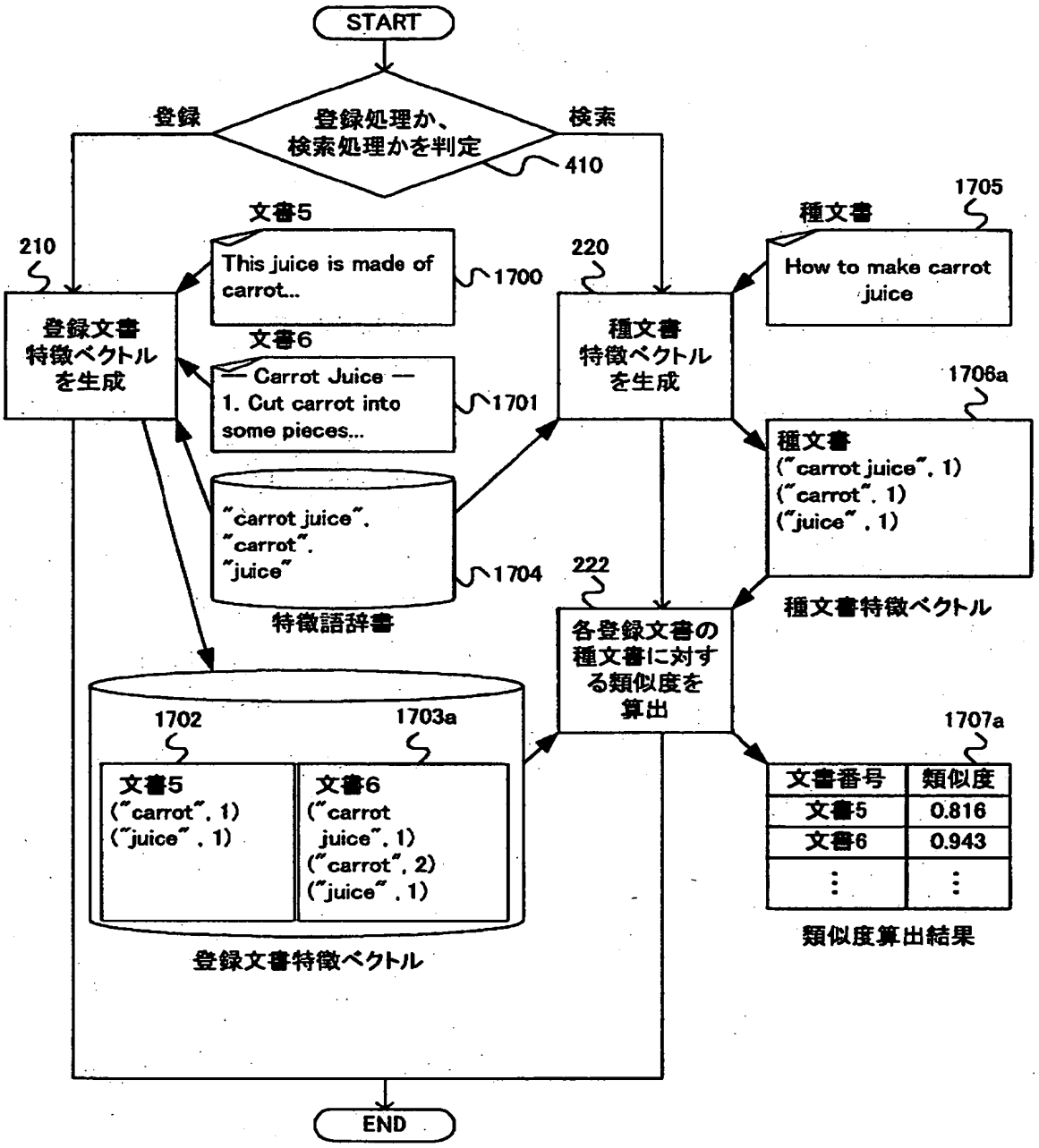
【図１１】

図11


START
登録処理か、検索処理かを判定
登録
検索
410
文書5
This juice is made of carrot...
1700
文書6
— Carrot Juice —
1. Cut carrot into some pieces...
1701
210
登録文書特徴ベクトルを生成
220
種文書特徴ベクトルを生成
種文書
How to make carrot juice
1705
1706
種文書
("carrot juice", 1)
"carrot juice",
"carrot",
"juice"
1704
特徴語辞書
222
各登録文書の種文書に対する類似度を算出
種文書特徴ベクトル
1702
1703
文書5
("carrot", 1)
("juice", 1)
文書6
("carrot juice", 1)
("carrot", 1)
登録文書特徴ベクトル
1707
文書番号
類似度
文書5
0.000
文書6
0.710
類似度算出結果
END

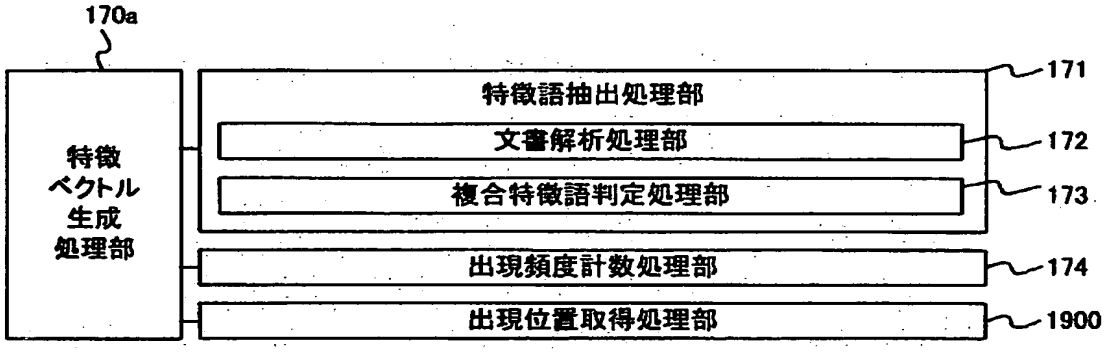
【図１２】

図12


START
登録処理か、検索処理かを判定
登録
検索
410
文書5
This juice is made of carrot...
1700
210
登録文書特徴ベクトルを生成
文書6
— Carrot Juice —
1. Cut carrot into some pieces...
1701
"carrot juice", "carrot", "juice"
1704
特徴語辞書
1702
1703a
文書5
("carrot", 1)
("juice", 1)
文書6
("carrot juice", 1)
("carrot", 2)
("juice", 1)
登録文書特徴ベクトル
220
種文書特徴ベクトルを生成
1705
種文書
How to make carrot juice
1706a
種文書
("carrot juice", 1)
("carrot", 1)
("juice", 1)
種文書特徴ベクトル
222
各登録文書の種文書に対する類似度を算出
1707a
文書番号
類似度
文書5
0.816
文書6
0.943
類似度算出結果
END

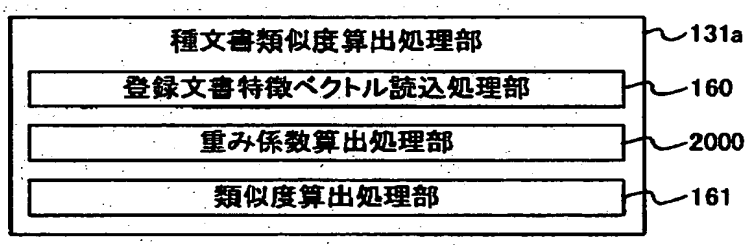
【図１３】

図13



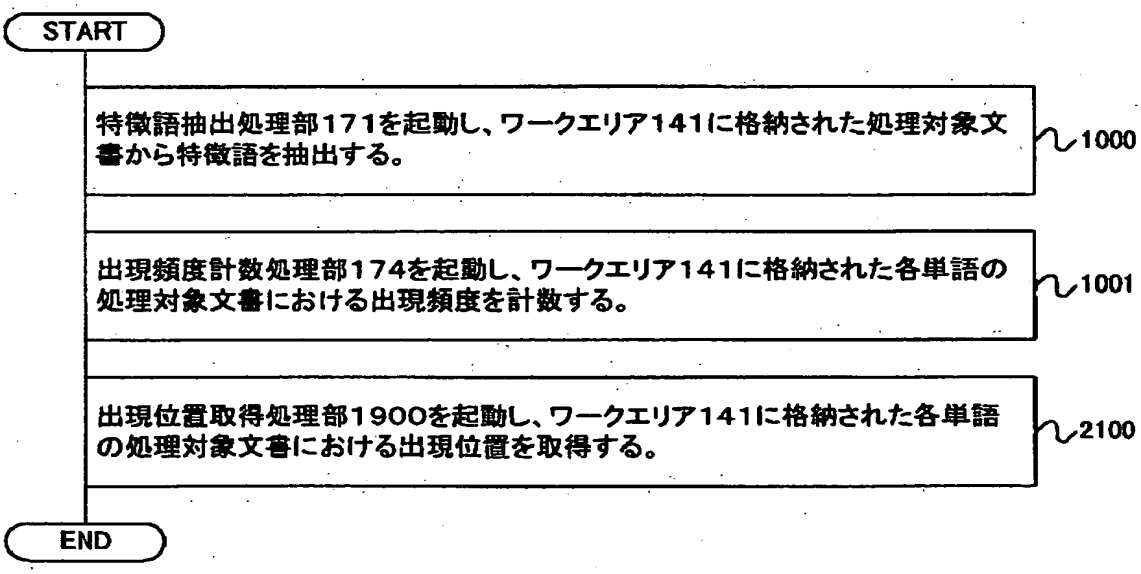
【図１４】

図14



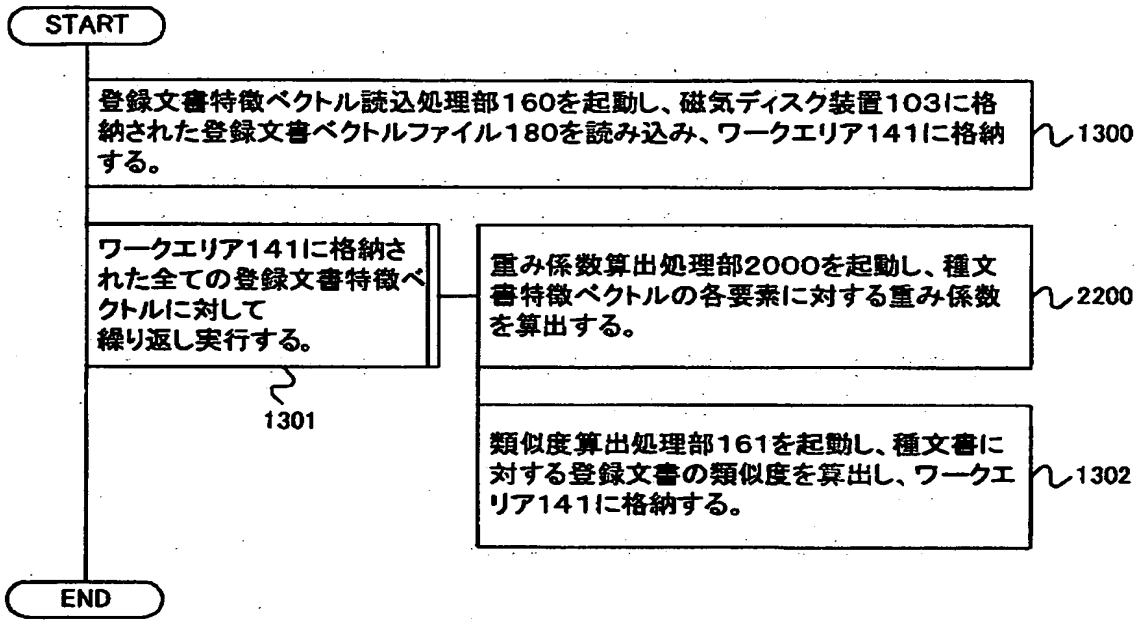
【図１５】

図15



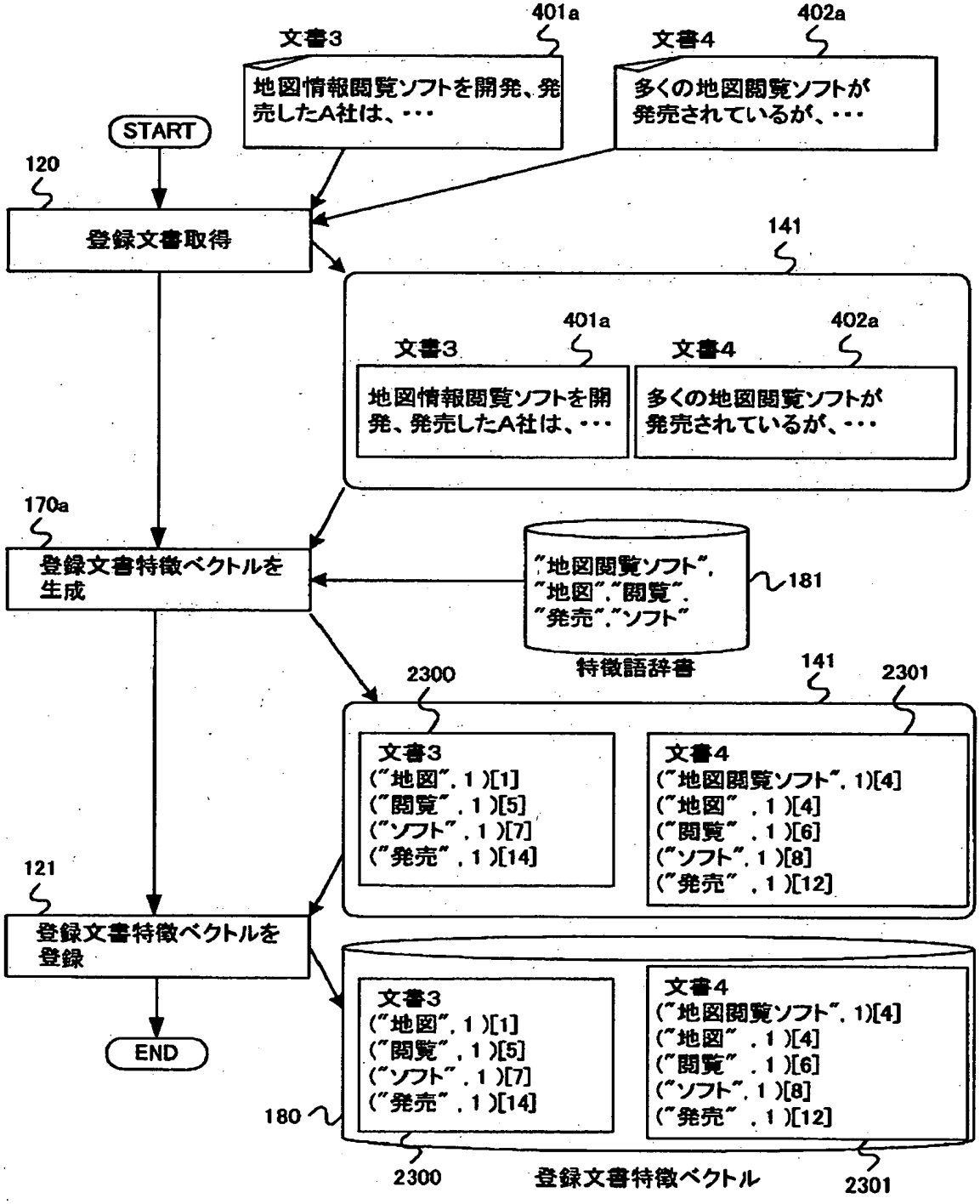
【図１６】

図16


START
登録文書特徴ベクトル読込処理部160を起動し、磁気ディスク装置103に格納された登録文書ベクトルファイル180を読み込み、ワークエリア141に格納する。
1300
ワークエリア141に格納された全ての登録文書特徴ベクトルに対して繰り返し実行する。
1301
重み係数算出処理部2000を起動し、種文書特徴ベクトルの各要素に対する重み係数を算出する。
2200
類似度算出処理部161を起動し、種文書に対する登録文書の類似度を算出し、ワークエリア141に格納する。
1302
END

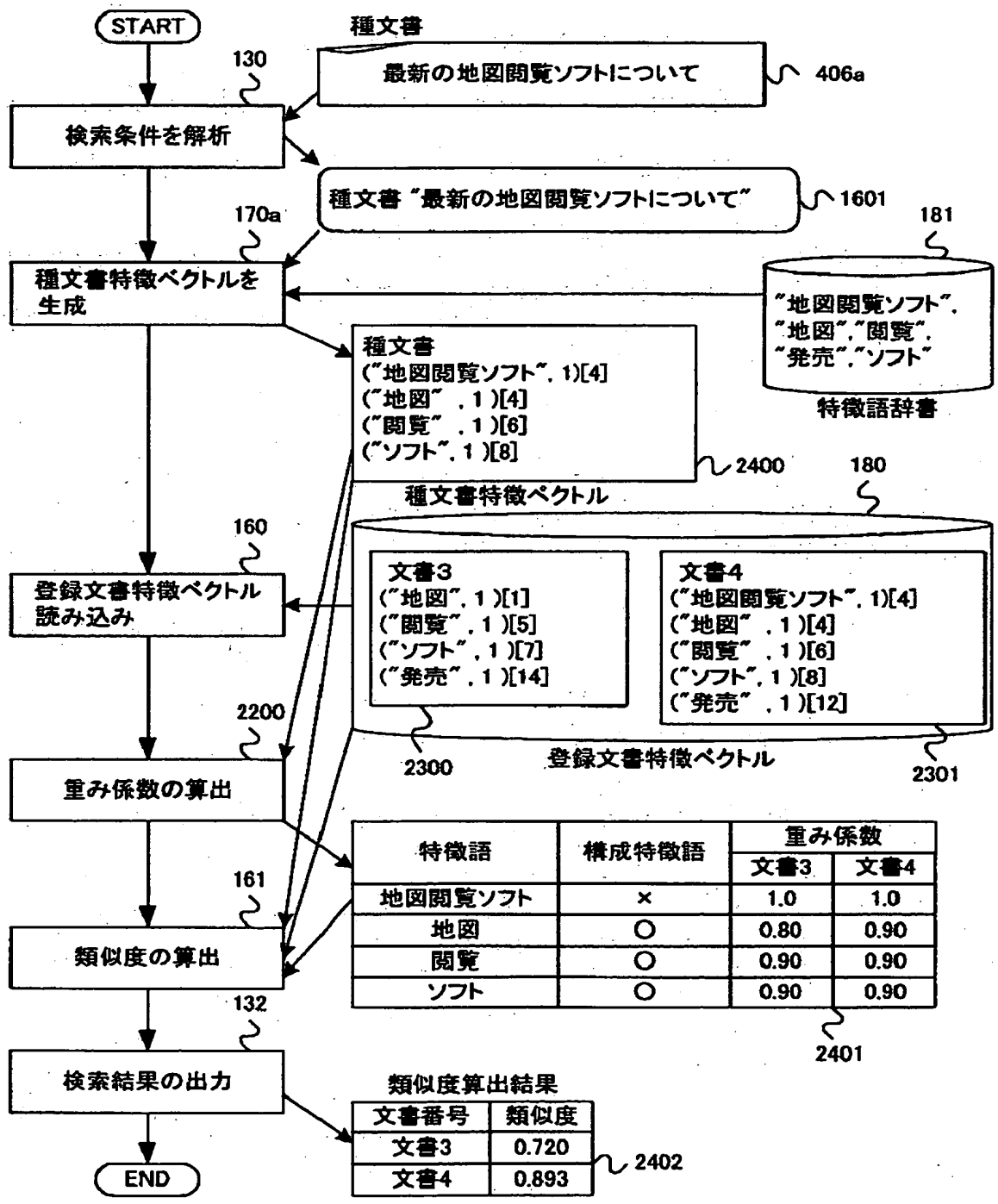
【図１７】

図17


401a
402a
文書3
文書4
地図情報閲覧ソフトを開発、発売したA社は、・・・
多くの地図閲覧ソフトが発売されているが、・・・
START
120
登録文書取得
141
401a
402a
文書3
文書4
地図情報閲覧ソフトを開発、発売したA社は、・・・
多くの地図閲覧ソフトが発売されているが、・・・
170a
登録文書特徴ベクトルを生成
"地図閲覧ソフト"、"地図"、"閲覧"、"発売"、"ソフト"
181
特徴語辞書
2300
141
2301
文書3
("地図", 1)[1]
("閲覧", 1)[5]
("ソフト", 1)[7]
("発売", 1)[14]
文書4
("地図閲覧ソフト", 1)[4]
("地図", 1)[4]
("閲覧", 1)[6]
("ソフト", 1)[8]
("発売", 1)[12]
121
登録文書特徴ベクトルを登録
END
180
文書3
("地図", 1)[1]
("閲覧", 1)[5]
("ソフト", 1)[7]
("発売", 1)[14]
文書4
("地図閲覧ソフト", 1)[4]
("地図", 1)[4]
("閲覧", 1)[6]
("ソフト", 1)[8]
("発売", 1)[12]
2300
登録文書特徴ベクトル
2301

【図１８】

図18

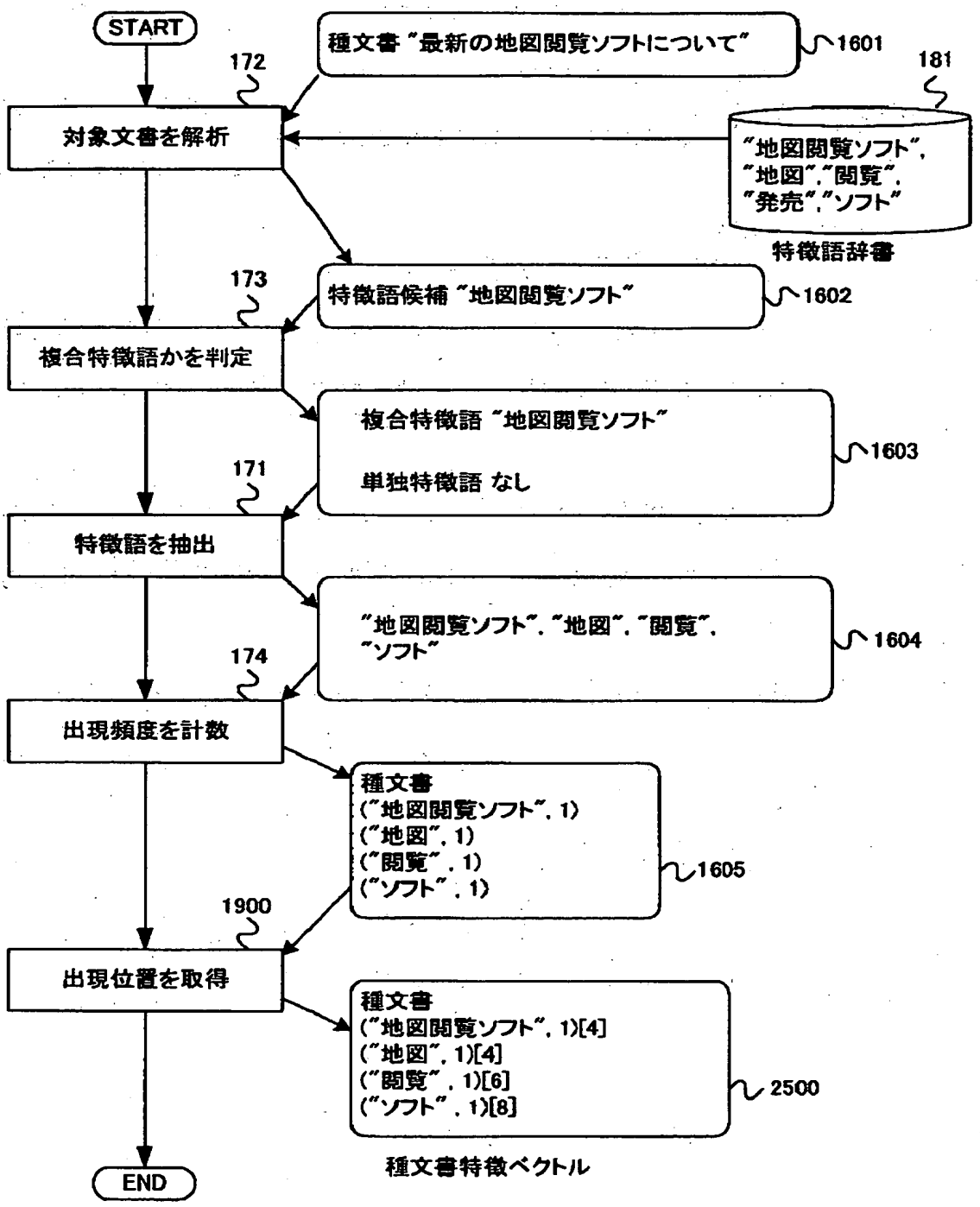


特徴語	構成特徴語	重み係数	
		文書3	文書4
地図閲覧ソフト	×	1.0	1.0
地図	○	0.80	0.90
閲覧	○	0.90	0.90
ソフト	○	0.90	0.90

2401

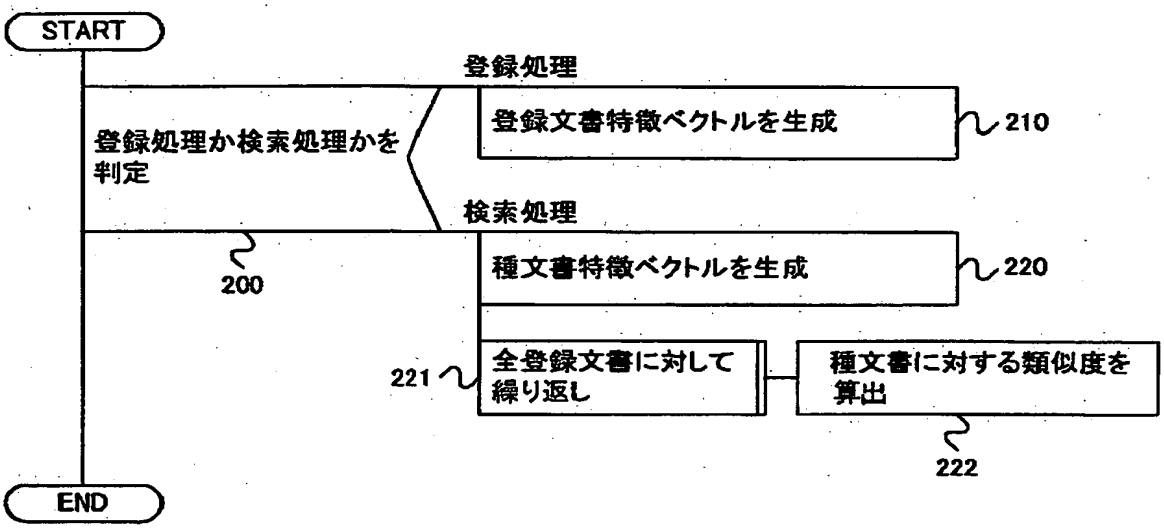
【図１９】

図19


START
種文書 "最新の地図閲覧ソフトについて"
1601
172
181
対象文書を解析
"地図閲覧ソフト",
"地図","閲覧",
"発売","ソフト"
特徴語辞書
173
特徴語候補 "地図閲覧ソフト"
1602
複合特徴語かを判定
複合特徴語 "地図閲覧ソフト"
単独特徴語 なし
1603
171
特徴語を抽出
"地図閲覧ソフト","地図","閲覧",
"ソフト"
1604
174
出現頻度を計数
種文書
("地図閲覧ソフト", 1)
("地図", 1)
("閲覧", 1)
("ソフト", 1)
1605
1900
出現位置を取得
種文書
("地図閲覧ソフト", 1)[4]
("地図", 1)[4]
("閲覧", 1)[6]
("ソフト", 1)[8]
2500
種文書特徴ベクトル
END

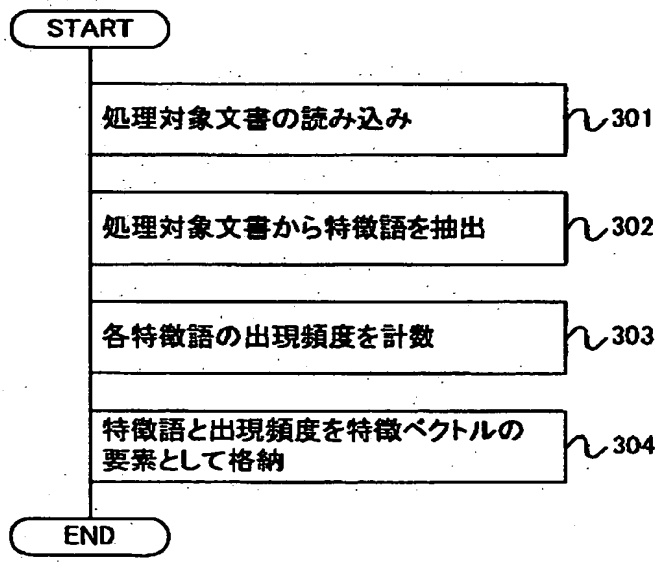
【図２０】

図20



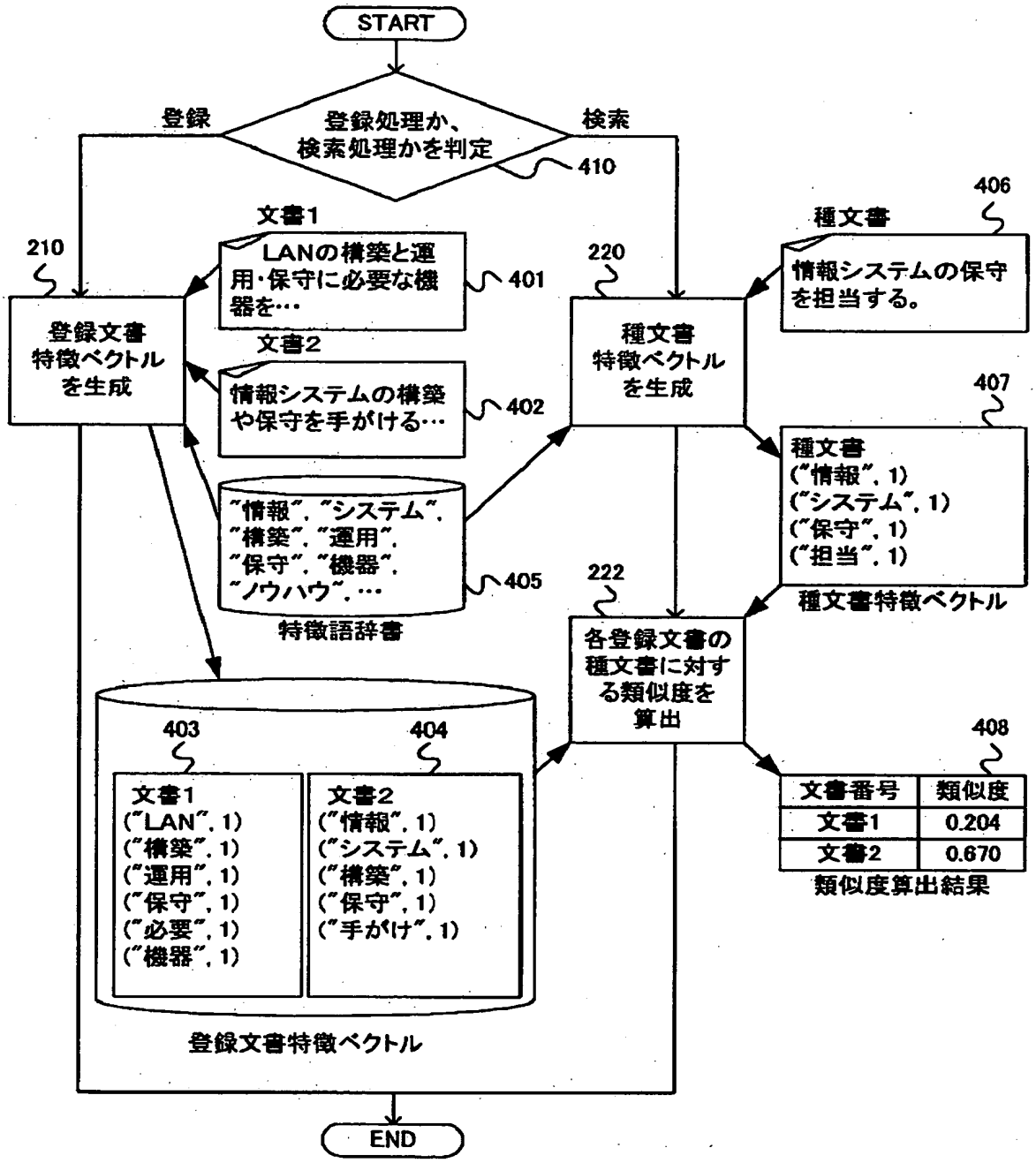
【図２１】

図21



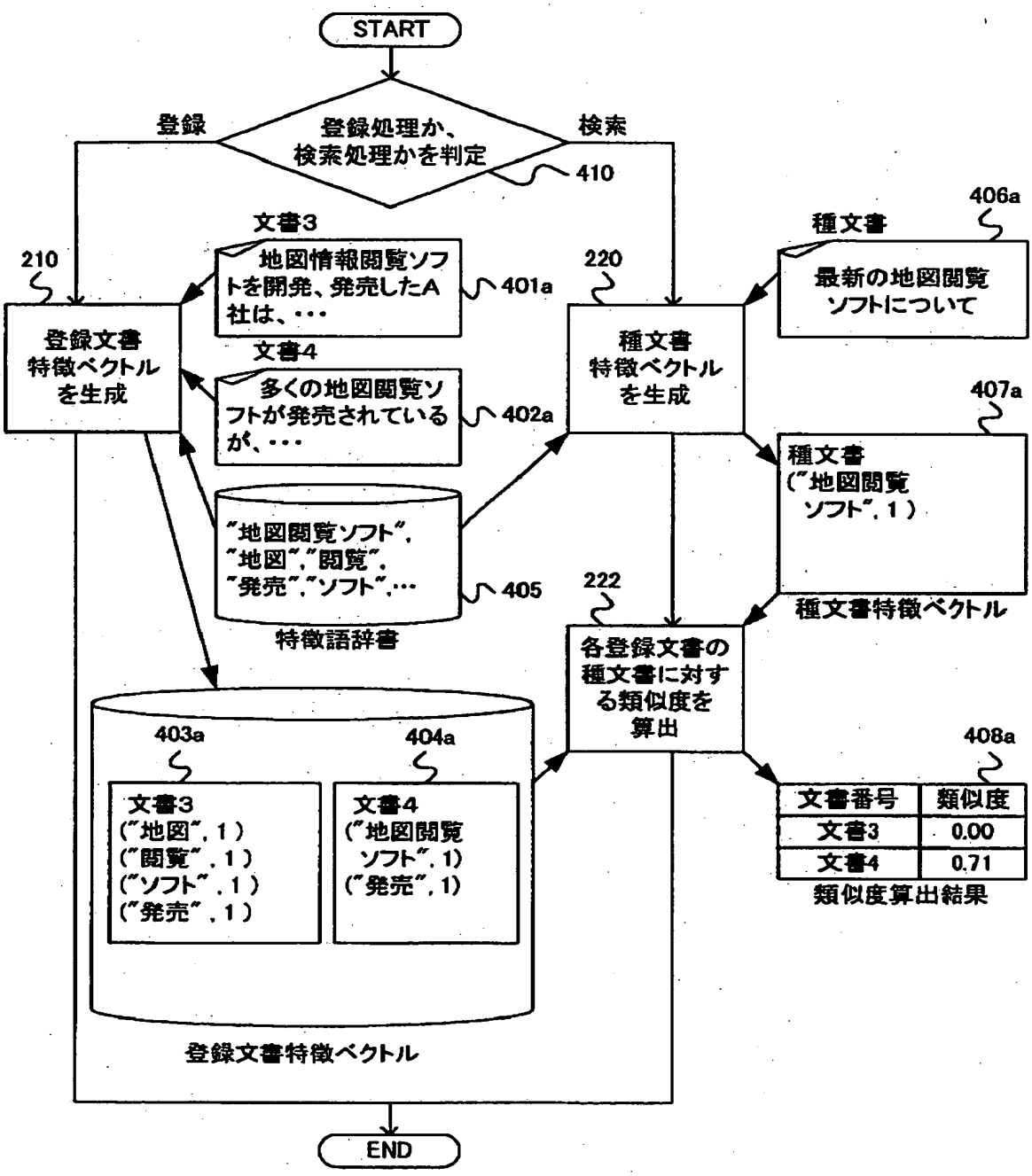
【図２２】

図22


START
登録処理か、検索処理かを判定
登録
検索
410
文書1
LANの構築と運用・保守に必要な機器を…
401
文書2
情報システムの構築や保守を手がける…
402
210
登録文書特徴ベクトルを生成
220
種文書特徴ベクトルを生成
406
種文書
情報システムの保守を担当する。
407
種文書
("情報", 1)
("システム", 1)
("保守", 1)
("担当", 1)
種文書特徴ベクトル
"情報", "システム", "構築", "運用", "保守", "機器", "ノウハウ", …
405
特徴語辞書
222
各登録文書の種文書に対する類似度を算出
403
404
文書1
("LAN", 1)
("構築", 1)
("運用", 1)
("保守", 1)
("必要", 1)
("機器", 1)
文書2
("情報", 1)
("システム", 1)
("構築", 1)
("保守", 1)
("手がけ", 1)
登録文書特徴ベクトル
408
文書番号 | 類似度
文書1 | 0.204
文書2 | 0.670
類似度算出結果
END

【図２３】

図23


START
登録処理か、検索処理かを判定
登録
検索
410
文書3
地図情報閲覧ソフトを開発、発売したA社は、・・・
401a
文書4
多くの地図閲覧ソフトが発売されているが、・・・
402a
210
登録文書特徴ベクトルを生成
220
種文書特徴ベクトルを生成
406a
種文書
最新の地図閲覧ソフトについて
407a
種文書（"地図閲覧ソフト"，1）
種文書特徴ベクトル
"地図閲覧ソフト"，"地図"，"閲覧"，"発売"，"ソフト"，・・・
405
特徴語辞書
222
各登録文書の種文書に対する類似度を算出
403a
文書3
("地図"，1)
("閲覧"，1)
("ソフト"，1)
("発売"，1)
404a
文書4
("地図閲覧ソフト"，1)
("発売"，1)
登録文書特徴ベクトル
408a
文書番号 | 類似度
文書3 | 0.00
文書4 | 0.71
類似度算出結果
END

【書類名】　　要約書

【要約】

【課題】　検索漏れの無い高精度な類似文書検索を実現し、内容が特に関連した文書を精度良く検索することが可能な技術を提供する。

【解決手段】　指定された文書と類似する文書を検索する類似文書検索方法において、所望の検索内容を含んだ種文書から特徴語の候補となる特徴語候補を抽出するステップと、前記抽出された特徴語候補が複数の特徴語で構成された複合特徴語である場合に当該特徴語候補から複合特徴語及びその複合特徴語を構成する構成特徴語を当該種文書の特徴語として抽出するステップと、前記抽出された種文書の特徴語と登録文書の特徴語との間の類似度を算出するステップと、前記算出された類似度算出結果を検索結果として出力するステップとを有するものである。

【選択図】　　図１

認定・付加情報

特許出願の番号	特願２００１－１２８９３４
受付番号	５０１００６１６２３１
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１３年　５月　２日

＜認定情報・付加情報＞

【提出日】　平成13年 4月26日

次頁無

出 願 人 履 歴 情 報

識別番号　〔０００００５１０８〕

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所